An exploratory performance assessment of the *CHIMERE* model (version 2017r4) for the northwestern Iberian Peninsula and the summer season

Swen Brands[1,2], Guillermo Fernández-García[1], Marta García Vivanco[3], Marcos Tesouro Montecelo[1], Nuria Gallego Fernández[4], Anthony David Saunders Estévez[2,4], Pablo Enrique Carracedo García[1], Anabela Neto Venâncio[1,2], Pedro Melo Da Costa[1,2], Paula Costa Tomé[2,4], Cristina Otero[2,4], María Luz Macho[1], and Juan Taboada[1,2]

[1]MeteoGalicia - Consellería de Medio Ambiente, Territorio e Vivenda, Xunta de Galicia, Santiago de Compostela, Spain
[2]Tragsatec, Santiago de Compostela, Spain
[3]Centro de Investigaciones Energéticas, Medioambientables y Tecnológicas (CIEMAT), Madrid, Spain
[4]Servicio de Calidad del Aire - Consellería de Medio Ambiente, Territorio e Vivenda, Xunta de Galicia, Santiago de Compostela, Spain

Correspondence: Swen Brands (swen.brands@gmail.com)

Abstract. Here, the capability of the chemical weather forecasting model CHIMERE (version 2017r4) to reproduce surface ozone, particulate matter and nitrogen dioxide concentrations in complex terrain is investigated for the period from June 21 to August 21, 2018. The study area is the northwestern Iberian Peninsula, where both coastal and mountain climates can be found in direct vicinity and a large fraction of the land area is covered by forests. Driven by lateral boundary conditions from ECMWF's Composition Integrated Forecast System, anthropogenic emissions from two commonly used top-down inventories and meteorological data from the Weather Research and Forecasting Model, CHIMERE's performance with respect to observations is tested with a range of sensitivity experiments. We assess the effects of 1) an increase in horizontal resolution, 2) an increase in vertical resolution, 3) the use of distinct model chemistries and 4) the use of distinct anthropogenic emissions inventories, downscaling techniques and land use databases. In comparison with the older HTAP emission inventory downscaled with basic options, the updated and sophistically downscaled EMEP inventory only leads to partial model improvements and so does the computationally costly horizontal resolution increase. Model performance changes caused by the choice of distinct chemical mechanisms are not systematic either and rather depend on the considered anthropgenic emission configuration and pollutant. Albeit the results are thus heterogeneous in general terms, the model's response to a *vertical* resolution increase confined to the lower to middle troposphere is homogeneous in the sense of improving virtually all verification aspects. For our study region and the two aforementioned top-down emission inventories, we conclude that it is not necessary to run CHIMERE on a horizontal mesh much finer than the native grid of these inventories. A relatively coarse horizontal mesh combined with 20 model layers between 999 and 500 hPa is sufficient to yield balanced results. The chemical mechanism should be chosen as a function of the intended application.

1 Introduction

Motivated by the air quality legislation of the European Union (EU, 2008), many governmental air quality departments are currently demanding air quality forecasting schemes based on numerical models (Thunis et al., 2016), and the need for accurate and computationally efficient predictions in this field is perhaps greatest than ever before. For Europe as a whole, the most important real-time prediction system available to date is provided by the Copernicus Atmosphere Monitoring Service (Marécal et al., 2015), comprising an ensemble of currently seven chemical weather forecasting (CWF) models[1] run for the entire continent at a horizontal resolution of 0.1°. to 0.25° in longitude and 0.1° to 0.2° in latitude. In addition to this short-term prediction system, several large research initiatives have been issued during the last two decades in order to assess the *climatological* properties of atmospheric composition, including the detection of long-term trends resulting from emission reductions induced by the Convention on Long-range Transboundary Air Pollution (CLRTAP, 2019). The final aim of these efforts is to find model configurations, or ensembles thereof, that can be used as surrogates for real observations in order to assess whether emission reductions actually have lead, or would lead, to changes in the atmosphere's composition on climatological timescales (Vautard et al., 2006; Jonson et al., 2006; Colette et al., 2011; Wilson et al., 2012; Banzhaf et al., 2015; Colette et al., 2017; Im et al., 2018b, a; Vivanco et al., 2018; Theobald et al., 2019).

Complementary to these large-scale efforts, usually conducted with a single configuration of a given model (Bessagnet et al., 2016), small-scale sensitivity tests for particular models are still relevant since they can be run with more sophisticated model configurations than their large-scale counterparts and are therefore more interesting for regional prediction systems, such as those demanded by national or regional governments (Banzhaf et al., 2012; Beegum et al., 2016; Flamant et al., 2018). Further, following the concept of seamless prediction (Palmer et al., 2008), lessons learned from short-term prediction systems for relatively small geographical areas might as well be important for longer lead-times and larger areas.

Previous sensitivity studies have identified several *factors* influencing the models' capability to correctly reproduce observed values, hereafter referred to as "model performance" (Giorgi and Francisco, 2000; Chang and Hanna, 2004). Among these factors, the meteorological model used to drive the chemical model and the accuracy of the underlying emission datasets play a key role and have been assessed in a number of studies (Menut, 2008; Markakis et al., 2015; Colette et al., 2017; Otero et al., 2018; Vivanco et al., 2018). The resolution of the model mesh used to discretize the chemical reactions and atmospheric dynamics is also important and, when it is increased, a trade-off between potential performance gains and computational cost must been made in practice. In what concerns the *horizontal* resolution, performance gains have been reported up to a scale of approximately 12 km for a number of models, such as WRF-CHEM and CHIMERE (Valari and Menut, 2008; Schaap et al., 2015; Crippa et al., 2017). However, a further resolution increase does not guarantee further performance gains. Beyond the 12 km threshold, Misenis and Zhang (2010) reported heterogeneous results for WRF-CHEM that strongly depend on the considered time period. For the use of CHIMERE and focussing on surface O_3 concentrations, Valari and Menut (2008) even found a performance *loss* which they attributed to a noise increase in the emission fluxes and meteorological input data at higher resolutions. Regarding the role of *vertical* resolution, an increase therein has been found to improve the modelled

[1] see Kukkonen et al. (2012) for an overview of these models

particulate matter (PM) concentrations during desert dust events when using WRF-CHEM (Teixeira et al., 2016). CHIMERE's performance, however, was found to be only weakly affected by this kind of resolution increase (Menut et al., 2013a; Markakis et al., 2015).

Representing the number and complexity of the considered chemical reactions, several *chemistry mechanisms* are usually available for a given model and switching from one mechanism to another can also affect the model's performance (Balzarini et al., 2015; Karlický et al., 2017). In recent CHIMERE versions, the SAPRC-07A mechanism (hereafter: SAPRC) has been included as an alternative to the full or reduced versions of the Melchior mechanism (Carter, 2010; Mailler et al., 2017) but, to the authors' knowledge, related sensitivity tests are sparse to date.

A common limitation of small-scale sensitivity studies is that their conclusions, strictly speaking, only hold for the considered region, time period or season of the year. In this context, most of the aforementioned conclusions for CHIMERE (the model applied here) have been drawn for the *Île de France* region, which is densely populated, relatively flat and not directly influenced by sea-salt emissions. The model has been applied for a number of other regions but the map is still incomplete and sensitivity testing is not the main focus of the corresponding studies (Mazzeo et al., 2018; Menut et al., 2018; Monteiro et al., 2018; Brasseur et al., 2019; Deroubaix et al., 2019).

This is where the present study comes into play: For the two month period from June 21 to August 21, 2018 a series of 19 sensitivity tests has been run with CHIMERE over the *northwestern Iberian Peninsula*, a region characterized by forested mountain terrain, a complex coastline and the advection of sea-salt from the surrounding Atlantic Ocean, quite different from the *Île de France* region. The applied tests will quantify the effects arising from 1) an increase in model resolution (vertical and/or horizontal), 2) switching from one chemistry mechanism to another (full Melchior or SAPRC in this case) and 3) changing the applied anthropgenic emissions inventory, downscaling strategy and land use database. To this end, version 2017r4 of the CHIMERE model is used (Mailler et al., 2017) in combination with the HTAP v2.2 and EMEP emission inventories of the years 2010 and 2017 respectively (Janssens-Maenhout et al., 2015; EMEP/CEIP, 2019). Long-range transport events of e.g. ozone and its precursors or Saharan dust are accounted for by passing them through from a globel model at the lateral boundaries of the outer CHIMERE domain (Fig. 1a). This way, it is not necessary to run CHIMERE on a large domain covering all relevant remote emission sources (Bessagnet et al., 2017; Gama et al., 2020) which frees computational resources that are put into our region of interest instead. The global model data applied here for this purpose are from the operational forecasts run with the European Centre for Medium-Range Weather Forecasts (ECMWF) Composition Integrated Forecasting system (C-IFS)(Flemming et al., 2015).

In Section 2, the applied data, model configurations and verification measures are described. Results are presented in Section 3 and a discussion and some general conclusions are provided in Section 4.

2 Data and Methods

In this section, the meteorological input data and general characteristics of the CHIMERE experiments are depicted first (Section 2.1), followed by a description of the two applied emission inventories (Section 2.2) and individual model experiments

(Section 2.3). The in-situ station network used as reference for verification is introduced in Section 2.4. The section closes with a description of the verification measures used to estimate CHIMERE's performance for the applied experiments (see Section 2.5).

2.1 Meteorological Input and General Characteristics of the CHIMERE Experiments

The meteorological input data for the CHIMERE experiments is provided by the Weather Research and Forecasting (WRF) model version 3.5 (Skamarock et al., 2008), driven by Global Forecast System (GFS) forecasts initialized at 00 UTC (Caplan et al., 1997). WRF is run on three domains: a continental-scale domain having a resolution of 36km, followed by a regional domain covering southwestern Europe at a resolution of 12km and, finally, a 4km domain covering our study region, the northwestern Iberian Peninsula. For these domains, WRF is executed with a minimum time step of 216, 72 and 24 seconds and a maximum time step of 360, 180 and 60 seconds, respectively. All domains comprise 33 vertical layers with a model top at 10 hPa. A detailed overview of the WRF physics can be found in Table 1. In this configuration, WRF has been run for now more than a decade at the meteorological office of the Galician government (MeteoGalicia) in order to provide real-time meteorological forecasts for the northwestern Iberian Peninsula. It is able to simulate orographic and coastal effects on the local weather reasonably well, which is illustrated in supplementary Figure 1 for a typical summertime heat day (August 5th, 2018).

With this meteorological input, version 2017r4 of the CHIMERE model is run on two domains: a coarse domain having a horizontal resolution of $0.15° \times 0.15°$ ($longitude \times latitude$), and a fine domain, nested into the former, having a resolution of $0.05° \times 0.04°$ (see Figure 1a). Note that the terms "coarse" and "fine" shall hereafter refer to the CHIMERE domains, not the WRF domains, if not otherwise stated. Biogenic emissions comprising VOCs and NO are from the MEGAN model version 2.04 (Guenther et al., 2006) and mineral dust emissions within the CHIMERE domains are calculated on the basis of the United States Geological Survey (USGS) land use dataset (Loveland et al., 2000). The Alfaro and Gomes (2001) saltation and sandblasting scheme, optimized by Menut et al. (2005), and the surface wind threshold described in Shao and Lu (2000) are used throughout all experiments. The effect of soil moisture on dust emissions (Fécan et al., 1998) is activated and so are sea-salt emissions. Vertical advection is achieved by the upwind scheme, horizontal advection by the more complex van Leer (1979) scheme. Carbonaceous species as well as the interaction between aerosols and gases are taken into account by the model and the number of Gauss-Seidel iterations is set to 3 because the model occasionally develops unrealistic waves with lower numbers. Wind speed reduction in urban areas (the so called "urban correction") is deactivated, and so is the resuspension process. A complete list of the internal CHIMERE parameters common to all sensitivity experiments is provided in Table 2. For a full description of these parameters, the interested reader is referred to the CHIMERE user manual available at http://www.lmd.polytechnique.fr/chimere.

Along the lateral boundaries of the coarse domain, the concentrations of the chemical species required by CHIMERE are provided by three-hourly forecasts of the ECMWF Composition Integrated Forecasting System (C-IFS) initialized at 00 UTC (Flemming et al., 2015). This global model comprises 60 vertical levels and has a horizontal resolution of $\approx 80km$. In case a chemical species required by CHIMERE is not provided by C-IFS, the monthly climatological mean values from the MACC

reanalysis (Inness et al., 2013) are used instead. As an exception, sea-salt aerosols from MACC are applied albeit they are also available from C-IFS because the latter system was found to overestimate the corresponding concentrations in our study region. This bias is of minor importance for the summer season considered here, but would lead to an overestimation of the PM concentrations in the other, stormier seasons of the year. Similarly, the applied dust aerosols from C-IFS are scaled by a factor of 0.6 in order to compensate the positive bias observed during the two Saharan dust events occurring in the time period considered here. For all other chemical species from C-IFS, a scaling factor of 1 (i.e. no scaling) is used. The fact that the chemical and physical boundary conditions for our CHIMERE forecasts come from different prediction systems is assumed to be of minor importance for the short lead-times analyzed here (27 hours from initialization at the utmost).

To eliminate unwanted effects related to the spin-up, the daily WRF forecasts are initialized with the Digital Filtering Initialization (DFI) technique (Skamarock et al., 2008) and the first 3 integration hours are *not* used as meteorological input to CHIMERE. Consequently, CHIMERE is initialized on 03 UTC, using initial conditions from the model execution of the previous day, and is then integrated until 03 UTC of the following day to complete one forecast day. This procedure is repeated for each day from June 20, 2018 to August 21, 2018 and the resulting model output is then concatenated to form time series covering the entire time period. The verification against surface observations as described in Section 2.5 begins on June 21st 03 UTC, so CHIMERE is permitted to spin-up during the first 24 hours of the integration.

2.2 Anthropogenic Emission Inventories, Land Use Databases and Postprocessing

To assess CHIMERE's *combined* sensitivity to changes in the anthropogenic emissions, downscaling strategy and land use database, two distinct inventories and postprocessing techniques were selected: The EMEP dataset for the year 2017 on the one hand (EMEP/CEIP, 2019) and the HTAP v2.2 dataset for the year 2010 on the other (Janssens-Maenhout et al., 2015), both provided on a regular $0.1^\circ \times 0.1^\circ$ latitude-longitude grid. To disaggregate the raw data from these inventories, the publicly available program emiSURF shipped with the CHIMERE source code was used (Mailler et al., 2017), which was here modified to process EMEP data on the recently published $0.1^\circ \times 0.1^\circ$ grid. Spatial disaggregation is achieved by downscaling the emissions from their native grid to an auxiliary high-resolution grid at 1 km, followed by an upscaling to the two target domains displayed in Figure 1a. In this downscaling step, different proxies can be used to redistribute the raw emission data on the subgrid scale, among which land use categories are the standard option of the emiSURF program.

To downscale the raw HTAP v2.2 emissions, land use categories from the USGS dataset were used as the only proxy except for the "population downscaling" experiment, for which population density was used as an additional proxy (Gallego, 2010). The latter type of downscaling affects the NO_2 and particulate matter emissions from SNAP sector 2, originating mainly from domestic fuel burning (Mailler et al., 2017).

To spatially regrid the EMEP inventory, road traffic density and the locations of large point sources were used *in addition* to population density and land use categories, the latter provided by the GlobCover dataset in this case (Bicheron et al., 2011). The road traffic proxy affects the magnitude and allocation of the NO_2 emissions caused by this kind of activity whereas the locations of large point sources are used to re-allocate the corresponding emissions on the subgrid scale. The temporal disaggregation of the raw anthropogenic emission data to the timescale required by CHIMERE was accomplished by using

seasonal, weekly and hourly profiles for each pollutant and activity sector, which are implemented in the standard CHIMERE pre-processors (Menut et al., 2012; Mailler et al., 2017).

The above explained large differences between the spatial downscaling procedures of the two emission inventories were applied *intentionally* to assess CHIMERE's performance for the use of 1)an up-to-date and sophistically downscaled inventory (EMEP) versus 2) an older inventory downscaled with basic parameters (HTAP v2.2). For ease of understanding, these will hereafter be referred to as "emission configuration 1" and "emission configuration 2" respectively.

2.3 Specific Configuration of the Sensitivity Tests

To explore the influence of *vertical* resolution on model performance, 10 layer experiments are compared to 20 layer experiments, the lowermost layer being located at 999 hPa and the uppermost at 500 hPa in all cases (see Figure 1c+d). Thus, an increase in vertical resolution refers to a refinement in the lower to middle troposphere. An extension of the model top to, e.g., 200 hPa has been proposed in previous studies since some dust intrusions may extend to pressure levels above 500 hPa (Bessagnet et al., 2017). However, by design of our experiments, most of the dust intrusions' trajectory is simulated by the global atmospheric composition model providing the lateral boundary conditions (C-IFS) rather than internally simulated by CHIMERE and, therefore, elevating the model top is assumed to be of minor importance here.

The effect of an increase in *horizontal* resolution is tested by comparing the model output obtained with the coarse resolution domain with that of the fine resolution domain nested therein (see Figure 1a,e and f). In all but one fine resolution experiment (the "coarse meteorology" experiment defined below) the horizontal resolution increase is undertaken in *both* CHIMERE and WRF, meaning that the combined effect is assessed.

Version 2017r4 of the CHIMERE model offers the possibility to use three distinct "chemical mechanisms" describing the gas-phase chemistry considered by CHIMERE. The "full Melchior" mechanism consists of 300 reactions and 80 gaseous species and is the most complete but also the most computationally demanding of three. This is why a reduced version with 120 reactions and 40 species, the so called "reduced Melchior" or "Melchior 2" mechanism, is available as well. From version 2016a onwards, the SAPRC mechanism is implemented as the third mechanism (Carter, 2010), offering a chlorine chemistry not considered in any of the two Melchior mechanisms (Mailler et al., 2017). With 72 gaseous species and 218 chemical reactions, SAPRC's complexity and computational costs are somewhat lower than for full Melchior, but superior to reduced Melchior. For the European summer 2015, *reduced* Melchior and SARPC have been compared in Menut et al. (2013b), who found large differences in the composition of organic nitrogen between the two which could potentially influence the spatial distribution of ozone production. They also found that the systematic overestimation of surface ozone reported in many CHIMERE studies is slightly less a problem when using SAPRC. In the present study, however, the *full* version of the Melchior mechanism is applied instead of the reduced one, meaning that the aforementioned findings might not hold here.

All the aforementioned model configurations, comprising 2 horizontal and 2 vertical resolution setups as well as 2 chemical mechanisms, are run separately with emissions configuration 1 and 2 as defined in Section 2.2.

Finally, three additional sensitivity tests are applied with constant anthropogenic emissions (HTAP), horizontal and vertical resolution (fine mesh, 20 layers), chemistry mechanism (full Melchior) and land use database (USGS). First, the effects of using

the population proxy for downscaling the raw HTAP emissions are explored in what is called the “Population downscaling” experiment (FM20H-P) hereafter. Then, the fine horizontal CHIMERE mesh is run with the *coarse* WRF mesh in the “Coarse meteorology” experiment (FM20H-C) in order to see whether low resolution meteorological input deteriorates CHIMERE’s performance. Finally, the effects of missing biogenic emissions are explored by intentionally turning them off in the "No biogenic emissions" experiment (FM20H-N).

For further reading, it is helpful to understand the rationale behind the abbreviations used for the distinct experiments. The first letter of a given abbreviation refers to the horizontal resolution of the CHIMERE mesh (C = Coarse or F = Fine), the second letter to the applied chemical mechanism (M = Melchior or S = SAPRC) and the following number to the vertical model levels used in the experiment (10 or 20). The third letter then points to the applied anthropogenic emission configuration (E = EMEP or H = HTAP) and the optional fourth letter separated by a hyphen to one of the three specific experiments described above (P = Population downscaling, C = Coarse meteorology, N = No biogenic emissions).

An overview of all applied sensitivity tests is provided in Table 3. In the last column, the computational costs for a typical summertime heat day (August 5th, 2018) are listed for the emission configuration 1 experiments. The run times of the respective configuration 2 experiments are in very close agreement (e.g. for CS10H and CS10E) but cannot be exactly stated since they were unfortunately not saved.

2.4 The Air Quality Monitoring Network in Northwestern Spain (Galicia)

The Galician air quality monitoring network comprises a total of 46 stations which, as a function of the main pollution source or the lack thereof, can be grouped into background, industrial and traffic sites (see Figure 1b). Currently, 14 stations are directly maintained by the Galician regional government (Xunta de Galicia). The remaining 32 stations are maintained by industrial companies which are supervised by the government in order to assure the same measurement standards, specified in the national UNE-EN norm.

The quality control of the corresponding data is accomplished *manually* by trained technical staff of the regional government, i.e. is centralized in one institution. First, outlier values are detected by comparing a suspicious value to the typical time series behaviour at the considered site and at the surrounding sites. Once the outlier is detected, its validity is determined taking into account inter-variable relationships, potential power breakdowns, calibration errors, damages and changes in the topographic features surrounding the station. This way, a quality controlled observational dataset has been developed which, at some locations, is now nearly a decade long. This dataset serves as reference for model verification.

2.5 Applied Verification Measures

Here, the *temporal* agreement between the modelled and observed time series is measured in terms of the Pearson correlation coefficient (R), the percentage bias (see Equation 1), and the standard deviation ratio (see Equation 2):

$$BIAS = \frac{\overline{m} - \overline{o}}{\overline{o}} \times 100 \tag{1}$$

$$RATIO = \frac{\sigma_m}{\sigma_o} \tag{2}$$

, where $\overline{m}$, $\overline{o}$, σ_m and σ_o are the modelled and observed values for the temporal mean and standard deviation, respectively.

Note that the chosen verification measures are complementary to each other since they cover different time series aspects. Namely, BIAS and RATIO measure the model's capacity to reproduce the observed temporal mean and dispersion whereas R looks at the similarity in day-to-day variability irrespective of errors in the mean and dispersion. The perfect scores for BIAS, RATIO and R are 0, 1 and 1, respectively.

In addition, the mean absolute error (MAE) is a good measure of overall performance, and is here applied as a skill score (mean absolute error skill score, MAESS), i.e. as percentage deviation from the error of a reference experiment:

$$MAESS = \left(1 - \frac{MAE_i}{MAE_{ref}}\right) \times 100 \tag{3}$$

, where MAE_i is the error a specific experiment i and MAE_{ref} the error of the experiment CS10E, used as reference throughout the present study since it is the computationally least expensive experiment (see Table 3). Positive values indicate performance gains, negative values performances losses with respect to the reference (Jolliffe and Stephenson, 2012). These verification measures are applied to hourly mean observations and hourly model data as provided by CHIMERE and, also, to the daily minimum and maximum values obtained from the former. All verification results are for the lowermost model layer whose upper limit is located at 999 hPa, i.e. roughly 10m above ground.

The aforementioned temporal verification scores are calculated separately for each station exceeding the 80% threshold of valid values and are then visualized either by overlay maps or boxplots. The centre line of each boxplot refers to the median value of the group of point-wise temporal verification results and the box to the interquartile range (IQR) of this group. The whiskers extend from the 25th percentile minus 1.5 $\times$ IQR at the lower end to the 75th percentile plus 1.5 $\times$ IQR at the upper end. Outlier verification results lying beyond these limits are not shown since their inclusion would blow up the scale of the figures and thus hamper their interpretability.

Apart from these temporal verification scores, the spatial bias (SBIAS, S = "spatial"), correlation coefficient (SR), standard deviation ratio (SRATIO) and mean absolute error (SMAE) were calculated on the point-wise temporal *mean* values in order to assess whether the spatial statistics of the *average* pollutant concentrations are captured by the model. Likewise, the same scores have been applied on the point-wise temporal *standard deviation* values to assess whether the model reproduces the spatial statistics of *temporal variablity*.

3 Results

3.1 Maximum Values

3.1.1 Temporal Mean and Standard Deviation

Fig. 2 shows the temporal *mean* values of the daily *maximum* concentrations seen in observations (the dots) plotted on the respective model value (the underlying pattern) for the 4 experiments driven with emission configuration 1 and the chemical mechanism SAPRC (CS10E, CS20E, FS10E and FS20E, row 1-4). Rows are ordered so that the first pair refers to the coarse horizontal mesh and the second pair to the fine one. Further, the 10 and 20 vertical layer experiments are placed on top of each other to assess the effects of an increase in vertical resolution. In the fifth row, the fourth experiment (fine horizontal resolution, 20 layers) is replicated with emission configuration 2 to show the effects of a combined change in the choice of the anthroponenic emission inventory (from EMEP to HTAP), downscaling technique (from land use, populatuion and traffic downscaling to land use downscaling only) and land use database (from Globcvoer to USGS). The spatial bias (SBIAS, in ug/m3), correlation coefficient (SR), standard deviation ratio (SRATIO = $\sigma_{model}/\sigma_{obs}$) and mean absolute error (SMAE) of the modelled vs. observed temporal mean values for these experiments is provided in Table 4a.

An increase in *horizontal* resolution improves the model's performance for $PM2.5$ by reducing SBIAS and by bringing SRATIO closer to unity. For NO_2 and O_3, however, model performance either does not improve or clearly deteriorates. Most notably, SBIAS increases for both species and SRATIO does so for NO_2, eventually exceeding a value of 2 which means that the spatial dispersion of the modelled mean NO_2 maxima is more than twice the observed one. As will be shown below (see Section 3.1.2), these error increases are likely associated with the population downscaling technique used to disaggregate the raw EMEP emissions.

An increase in *vertical* resolution reduces SBIAS by up to 2.4 ug/m3 (i.e. 40%) for the mean O_3 values and by up to 0.9 ug/m3 (i.e. 80%) for the mean $PM2.5$ values. For the latter pollutant, vertical refinement is much more efficient when using the fine horizontal mesh, in which case SBIAS is clearly improved.

For the fine horizontal mesh and 20 vertical layers, a switch to emission configuration 2 (i.e. from FS20E to FS20H, compare rows 4 and 5) translates into an improvement of SRATIO for NO_2 and O_3 but to a worsening for $PM2.5$. Also, results for FS20H are in closer agreement with CS20E than with FS20E, which points to the fact that the temporal *mean* daily maximum concentrations are more sensitive to the particular setup of the downscaling technique than to differences in the raw emission inventories.

In all considered experiments, the simulated mean O_3 concentrations are considerably higher over the sea than over land, which is in line with Terrenoire et al. (2015) and might be explained by reduced dry deposition and nighttime destruction by NO_2 over the sea resulting from a reduced surface roughness and NO_2 availability there (Davies et al., 1992; O'Hare and Wilby, 1995). Since this land-sea contrast is not seen in observations, the SR values for all experiments is essentially zero. This can be either explained by the lack of off-shore background observations (note that all available coastal sides are affected by

urban pollution) or by the fact that the reduced ozone destruction over the sea is less pronounced in the model than in the real world, translating into a positive bias there.

Figure 4 shows the temporal *standard deviation* of the daily *maximum* concentrations as seen in observations vs. those seen in the model, i.e. the model's capability to reproduce the observed temporal variability. The respective spatial verification results are provided by Table 4b. In general, CHIMERE is plagued by underdispersion, i.e. tends to underestimate the temporal variability of the daily maximum concentrations (SBIAS is negative). An increase in *horizontal* resolution alleviates this problem for $PM2.5$ and even leads to overdispersion for NO_2 (i.e. to a positive SBIAS), but does not noticeably alter the results for O_3. For $PM2.5$, SR is much improved when considering the fine horizontal mesh. Contrary to the findings for the temporal mean, temporal variability is more sensitive to a horizontal resolution increase than to a vertical resolution increase. Except for $PM2.5$, the impact of a switch in the emission configuration is less pronounced for the temporal standard deviation than for the aforementioned temporal mean (compare rows 4 and 5 in Fig. 3 and 4)

3.1.2 Full Temporal Verification

Fig. 5 shows the verification results of *all* applied experiments as ordered in Table 3 for the daily *maximum* NO_2 and O_3 concentrations. The perfect score for a given verification measure is indicated by a red vertical line. As can be seen from the figure, the NO_2 concentrations are generally underestimated by the model, except for the four emission configuration 1 experiments run on a high horizontal resolution (see Fig. 4a). Emission configuration 2 is plagued by larger median biases (see vertical orange lines within the boxes) than configuration 1 but has the advantage of a lower spatial spread in the results (see width of the boxes and whiskers). When applying a high horizontal resolution, this bias is reduced on average but the aforementioned spread is largely increased. While the effects of a vertical resolution increase and/or switch in the applied chemical mechanism are negligible, the effect of population dowscaling is considerable. Namely, the smallest median bias but largest spatial spread among all experiments is yielded if the raw HTAP emissions are disaggregated this way (see FM20H-P in Fig. 4a).

The structure of the verification results for the standard deviation ratio (see Equation 2 in Section 2.5) is in very close agreement with the structure found for the percentage bias and virtually identical lessons are learned (compare panel b with a in Fig. 4).

The model's capability to reproduce the temporal sequence of the observed anomalies, here measured with the Pearson correlation coefficient (R), is most improved by an increase in the horizontal resolution (see Fig. 4c). Emission configuration 1 yields systematically better results than configuration 2 (compare experiments ending on E with those ending on H in Fig. 4c). As opposed to the bias, the spatial spread of the *correlation coefficient* is larger for the coarse horizontal resolution than for the fine one, particularly if emission configuration 1 is used (compare the spread of the "C..." type experiments in Fig. 4c). The full Melchior mechanism yields slightly better correlation coefficients than SAPRC and so does the use of 20 instead of 20 vertical layers.

As indicated by Fig. 4d, the MAESS of the reference experiment CS10E is improved only by the CM20E experiment, meaning that the use of 20 vertical layers together with the full Melchior mechanism is sufficient to achieve optimal results for

this measure. A horizontal resolution increase is *not* necessary and is actually counterproductive if emission configuration 1 is used.

The inclusion of the population proxy in the downscaling procedure of the HTAP inventory leads to a sharp decrease in the spatial median MAESS and to the largest spatial spread among all experiments (see FM20H-P in Fig. 4d). In comparison, the use of coarse meteorological input data or removal of biogenic emissions has much smaller effects on the model's performance (compare FM20H-C and FM20H-N with FM20H in Fig. 4d)

As shown in Fig. 4e and f, CHIMERE overestimates the temporal mean and underestimates the temporal variability of the daily maximum O_3 concentrations. The effect of the performance factors is similar for each of the four applied verification measures. In general, the respective error is improved by a vertical resolution increase and by applying full Melchior instead of SAPRC, but is deteriorated or not improved when the horizontal resolution is increased. As an exception, SAPRC generally yields better correlation coefficients if the fine horizontal mesh is used (see 4g). Contrary to Menut et al. (2013b), average O_3 concentrations are larger for the SAPRC mechanism than for full Melchior. When considering MAESS, the emission configuration is the most influential factor on model performance, with configuration 1 clearly outperforming configuration 2 (see Fig. 4h). As was the case for maximum NO_2, 20 vertical layers yield better results than 10 layers and, for the use of emission configuration 1, the 20 layer setup performs nearly as well for the coarse horizontal mesh than for the fine one, meaning that the former is again preferable in case computational resources are limited (see last column in Table 3).

The full temporal verification results for the daily maximum $PM2.5$ and $PM10$ concentrations are displayed in Fig. 5. As shown in panels a+b and e+f, CHIMERE generally underestimates the temporal mean value and variability for both particle size fractions. The most important performance factor is the emission configuration, yielding smaller bias values with configuration 1 (see Fig. 5a+e) and better correlation coefficients with configuration 2, particularly for the fine particles (Fig. 5c+g). The effects of a horizontal resolution increase depend on the considered emission configuration and particle size fraction. Namely, configuration 1 improves the bias and standard deviation ratio for both size fractions (Fig. 5a+b and e+f) but has no effect on the correlation coefficient (Fig. 5c+g). Configuration 2, in turn, improves the correlation coefficient of the fine particles (Fig. 5c) but does not affect the bias nor the standard deviation ratio for any of the two particle size fractions (5a+b and e+f). A vertical resolution increase improves the bias for both particle sizes and, if a fine horizontal mesh is applied in addition, also the standard deviation ratio for the fine particles. The correlation coefficient, however, cannot be improved by a horizontal resolution increase and even deteriorates for some experiments (Fig. 5c+g). Regarding overall performance as measured by the MAESS (5d+h), SAPRC yields better results than full Melchior for nearly all experiments and both size fractions. The most robust skill increases are *again* obtained with 20 vertical layers, the coarse horizontal resolution, the SAPRC mechanism and emission configuration 1 (CS20E). Albeit the performance increase at individual stations may be much larger for other experiments, CS20E yields positive MAESS values at *all* stations and for *both* particle sizes. If the fine horizontal resolution is used (FS20E), the average performance improves for $PM10$ but deteriorates for $PM2.5$. FS20H and FM20H-P perform equally well than CS20E on average, but are characterized by a larger spatial spread in the results.

The population downscaling experiment outperforms its base experiment or is comparable to it for both particle sizes (compare FM20H-P with FM20H in all panels of Fig. 5). Using coarse resolution meteorological input does not noticeably affect the

results, except for a clear decrease in correlation for the fine particles (compare FM20H-C with FM20H in Fig. 5c). A lack of biogenic emission, however, largely enhances the bias (compare FM20H-N with FM20H in Fig. 5a+e), reduces the correlation (Fig. 5c+g) and worsens the overall performance as measured by the MAESS (Fig. 5d+h).

3.2 Minimum Values

3.2.1 Temporal Mean and Standard Deviation

Fig. 6 shows the temporal *mean* values of the daily *minimum* NO_2, O_3 and $PM2.5$ concentrations seen in observations (the dots), plotted on the respective model value (the underlying pattern) for the five experiments assessed in Section 3.1.1. The respective spatial verification results are provided in Table 4c. For NO_2 (Fig. 6, column 1), the model underestimates the temporal mean concentrations on average (SBIAS < 0) and underestimates their spatial dispersion (SRATIO < 1). The spatial pattern of the observed mean values is not well reproduced either (SR < 0.25 in Fig. 6a, d, g, and j). While the former two error types can be improved by augmenting the horizontal resolution (compare panels a+d with panels g+j in Fig. 6), the latter one can be reduced by using emission configuration 2 (compare panel j with m). Similar to the results for the maxima, using 20 instead of 10 vertical layers does not noticeably improve the result for the NO_2 *minima* either (compare Fig. 6a with d and g with j).

As for the maxima, the average *minimum* O_3 concentrations (Fig. 6, column 2) in are overestimated by the model (SBIAS $>$ 0). However, the spatial pattern of the observed values is generally well reproduced (RS $\geq$ 0.65) and so is the spatial dispersion if the coarse horizontal mesh is used (SRATIO $\approx$ 1). Using the fine horizontal mesh on the one hand reduces the bias but, on the other, inflates the spatial dispersion (SRATIO > 1, compare Fig. 6b with h and e with k). Results are improved when 20 instead of 10 vertical layers are used in combination with the *fine* horizontal mesh (compare panels h and k) and the bias largely increases when emission configuration 1 is applied (compare panels k and n).

The temporal mean daily $PM2.5$ minima (Fig. 6, column 3) are on average overestimated by the model (SBIAS > 0), their spatial dispersion is underestimated (SRATIO well below unity) and their spatial pattern not well reproduced (low values for SR). A *horizontal* resolution increase improves the spatial dispersion but deteriorates the spatial pattern and increases the bias, meaning that the negative effects prevail for this factor (compare Fig. 6 c with i and f with l). A *vertical* resolution increase generally has little effect on the model's performance unless the horizontal resolution is as well increased, in which case the bias increases for $PM2.5$ (compare c with with f and i with l). As for the maxima, results for FS20H are generally more similar to CS20E than to FS20E.

Fig. 7 and Table 4d show the respective verification results for the *temporal standard deviation* of the daily minimum concentrations. For NO_2 (Fig. 7, column 1), the model on average underestimates the temporal variability (SBIAS < 0) and the associated spatial dispersion (SRATIO well below unity). With SR values ranging in between 0.35 and 0.52 the model captures the spatial pattern of temporal variability to a certain degree. Results are insensitive to a *vertical* resolution increase (compare Fig. 7a with d and g with j) but systematically improve if the *horizontal* resolution is augmented (compare a with g and d with j). The temporal variability of the O_3 minima (Fig. 7, column 2) is on average well reproduced by the model (SBIAS

≈ 0). However, the associated spatial pattern is missed (SR ≈ 0) and the dispersion overestimated (SRATIO > 1). Neither a horizontal nor a vertical resolution increase noticeably improves these results. The temporal variability of the $PM2.5$ minima (Fig. 7, column 3) is also well reproduced on average and some skill is obtained for the respective spatial distribution. As for the NO_2 minima, the degree of spatial dispersion is as well underestimated for the $PM2.5$ minima (SRATIO < 1) and can be improved by a horizontal resolution increase (compare panels c with i and f with l). Results for FS20H closely agree with those for FS20E, except for generally lower O_3 and higher $PM2.5$ concentrations (compare the last two rows in Fig. 7).

3.2.2 Full Temporal Verification

Fig. 8 shows the full temporal verification results for the daily *minimum* NO_2 and O_3 concentrations. For a correct interpretation of the results, it is here important to note that the observed minimum concentrations in our study region are generally low and that average differences of only a few ug/m3 can translate into large *percentage* bias values.

As can be seen from Fig. 8a+b, the temporal mean and standard deviation of the daily minimum NO_2 concentrations are considerably underestimated at nearly all stations in any of the applied experiments. The spatial median values for BIAS and RATIO can be improved with a horizontal resolution increase and either emission configuration 1 (FS10E, FM10E, FS20E and FM20E) or configuration 2 plus population downscaling (FM20H-P), implying that this kind of downscaling is key at this point. However, improvements in the spatial median can only be achieved at the expense of a large increase in the spatial spread of the results, which is in line with the findings obtained for the NO_2 maxima (see Section 3.1.2). For the correlation coefficient (Fig. 8c), emission configuration 1 performs better than configuration 2, full Melchior better SAPRC and the coarse horizontal mesh better than the fine one. In comparison, an increase in vertical resolution from 10 to 20 layers is less efficient in improving the correlation. Coarse resolution meteorological input data and missing biogenic emissions both slightly worsen the model performance for all applied verification measures (compare FM20H-C and FM20H-N with FM20H in panels a, c, e and g). When considering the MAESS (Fig. 8d), the spatial median performance for the base experiment (CS10E) cannot be improved by any of the applied alternative experiments and the aforementioned growth in the results' spatial dispersion due to population downscaling can be clearly seen for FM20H-P.

Similar to the respective results for the maximum concentrations, daily *minimum* O_3 concentrations are also on average overestimated by the model (Fig. 8e) and the results for all verification measures can be improved by applying the full Melchior mechanism and 20 vertical layers (Fig. 8e to h). Contrary to the maxima, the spatial median performance for the O_3 minima can be generally further improved by applying a fine horizontal mesh, in which case the unwanted increase in spatial spread is less pronounced than for the maxima. The overall performance in terms of MAESS (Fig. 8h) is very satisfactory for the coarse horizontal resolution experiments run with 20 vertical layers (see CS20E and CM20E), which is in line with the results for the maxima. However, due to the aforemantioned relatively weak increase in the spatial spread of the results, the use of a fine mesh is more tentative for the O_3 minima than for the maxima; particularly if emission configuration 1 is applied (compare CS20E, CM20E with FS20E and FM20E in Fig. 8h and 4h). Coarse resolution meteorological input data or missing biogenic emissions both have negligible effects on the results. Population downscaling, however, leads to a systematic improvement (compare FM20H-C, FM20H-N and FM20H-P with FM20H in Fig. 8h).

The full temporal verification results for the $PM2.5$ and $PM10$ minima are displayed in Fig. 9. The model systematically overestimates the temporal mean $PM2.5$ concentrations and also tends to overestimate the temporal variability (Fig. 9a+b). Using 20 vertical layers instead of 10 enhances the correlation coefficient on the one hand but on the other generally increases the bias and shifts the standard deviation ratio to values larger than unity (except for moving from CS10E to CS20E, see Fig. 9a,b,c). A horizontal resolution increase has similar effects which are, however, larger in magnitude. Switching from SAPRC to full Melchior improves the results for all measures and nearly all experiments and overall performance gains as measured by MAESS are largest for this kind of switch (see panels a to d). When spatial median values are considered, the MAESS obtained with emission configuration 2 are systematically better than those obtained with configuration 1 (see Fig. 9d). However, the spatial spread in the MAESS is larger for configuration 2 than for configuration 1. In comparison with FM20H, overall performance deteriorates for the population downscaling experiment (see FM20H-P) and, even more so, for the coarse meteorological input experiment (see FM20H-C). Missing biogenic emissions improve the MAESS on average, but also increase the spatial spread (see FM20H-N). Notably, the performance increase of the CM10E experiment (with respect to the base experiment CS10E) is positive at *every* station, which is rarely the case in the present study. Hence, the coarse *horizontal* mesh is again a straightforward option which already yields optimal results with a simple 10-layer setup if the full Melchior mechanism is applied.

For the $PM10$ *minima*, emission configuration 2 yields smaller bias values and more favourable standard deviation ratios than configuration 1 (Fig. 9e+f), but weaker correlation coefficients (panel g). Using full Melchior instead of SAPRC and 20 instead of 10 vertical layers reduces the bias for all experiments, both factors being of roughly equal importance for this pollutant and temporal aggregation. Correlation coefficients are also improved, but only for the experiments run with emission configuration 1. If emission configuration 2 is used, SAPRC yields roughly the same correlation coefficients than full Melchior (Fig. 9g). The standard deviation ratios are systematically better for SAPRC than for full Melchior and for 20 instead of 10 layers if the fine horizontal mesh is chosen. Regarding MAESS (Fig. 9h), performance losses caused by population downscaling or coarse resolution meteorological input are less pronounced for the coarse particles than for the fine ones (compare FM20H-P and FM20H-C with FM20H in Fig. 9d+h). As for the fine particles, the "no biogenic emissions" experiment is also plagued by an increased spatial variability in the MAESS and, unlike the results for the fine particles, suffers a spatial average performance *decrease* if compared to its base experiment (compare boxes and median values for FM20H-N with FM20H in Fig. 9d+h). As expected, the modelled mean values are more realistic when biogenic emissions are taken into account (compare FM20H with FM20H-N in Fig. 9e). As for the fine particles, optimal results are obtained with the coarse horizontal mesh run with only 10 layers and the full Melchior mechanism (see CM10E in panel Fig. 9h). Albeit being second choice for the fine particles, emission configuration 1 is first choice for the coarse ones.

3.3 Verification Results per Pollution Source

Figure 10 shows the spatial median MAESS with reference to the base experiment CS10E for all locations (row 1) and separately for background, industry and traffic locations (rows 2 to 4). The first column refers to the results for daily maximum

concentrations, the second to hourly concentrations and the third to daily minimum concentrations, respectively. Improvement over the base experiment is indicated by green, worsening by red colour shadings.

As can bee seen from the predominantly red color shadings in the first two columns of Fig. 10, the base experiment CS10E already provides a good overall skill, difficult to exceed when considering daily maximum or hourly concentrations. Among all suggested model improvement factors, the use of 20 instead of 10 vertical layers yields the most balanced increases in spatial *median* performance irrespective of the applied chemical mechanism (see CS10E and CM20 in these columns). Switching from the coarse to high *horizontal* resolution leads to large performance increases for *particular* pollutants and/or station types, but only at the expense of performance decreases for the remaining species and sides and thus to unbalanced results.

Irrespective of the applied emission configuration and number of vertical layers, the best results for the *maximum and hourly* NO_2 *values* are obtained with a *coarse horizontal resolution*, except at traffic stations where the fine horizontal mesh yields better results if, importantly, emission configuration 2 is used *without* population downscaling (compare FS10E, FM10E, FS20E and FM20E in Figure 10j and k). At traffic and industry sides, the worst results for the NO_2 maxima and hourly data are obtained with the fine horizontal mesh and emission configuration 1 (relying on population and traffic downscaling) and with configuration 2 plus population downscaling (note the similarity between FS10E, FM10E, FS20E, FM20E and FM20H-P in Fig. 10a,b,g,h,j,k). Hence, this kind of downscaling is not advantageous in these cases.

For daily minimum NO_2, the coarse horizontal resolution is again the best choices, but only in combination with emission configuration 1 (see CS10E, CM10E, CS20E and CM20E in panels c, f, i and l). Using the coarse horizontal resolution with configuration 2 instead yields heterogeneous results, i.e. better results at industrial sides are contrasted by worse results at traffic sides (compare CS10H, CM10H, CS20H and CM20H in panel i with panel l).

For O_3, emission configuration 1 performs systematically better than configuration 2. Among the emission configuration 2 experiments, it is again the population downscaling experiment that most closely resembles the results from the configuration 1 experiments (compare experiments ending on "E" with FM20H-P). Importantly, using 20 instead of 10 vertical layers yields performance gains in virtually *any* case, i.e. irrespective of the applied emission configuration, horizontal mesh, chemical mechanism, temporal aggregation and pollution source type, and is consequently the most *robust* model improvement factor for surface O_3 concentrations assessed here. Second best in this context is the use of the full Melchior mechanism instead of SAPRC. Note also that the results for the maxima and hourly data are less similar to each other than for the remaining pollutants.

As opposed to the findings for O_3, emission configuration 2 is the better choice for $PM2.5$, particularly considering daily minimum concentrations at all kind of sides, as well as as maximum and hourly concentrations at industrial and traffic sides. The effects of a vertical resolution increase are heterogeneous. At background sides (see second row in Fig. 10 and also Supplementary Figure 2), results are improved for the daily maxima but deteriorate for the minima, with very little effects on the hourly concentrations. At industrial and traffic sides, however, results generally worsen for this factor. At background sides, SAPRC is generally superior to full Melchior whereas the opposite is found at industrial and traffic sides. As for O_3, a horizontal resolution increase is not advantageous for $PM2.5$ either, except for the daily minimum concentrations at industrial and traffic sides when using emission configuration 1.

Model sensitivity is generally lower for $PM10$ than for the other three pollutants. Largest performance gains are obtained for daily maximum concentrations, particularly at traffic sides, if the fine horizontal mesh is used in combination with 20 vertical layers and emission configuration 1 (see experiments FS20E and FM20E in panels a, d, g, and j). The same mesh, however, yields largest performance *losses* for minimum concentrations at background sides if emission configuration 2 is applied (see panel f). Albeit the differences are generally weak, the SAPRC mechanism is preferable for maximum and hourly concentrations whereas full Melchior is preferable for the minima.

Among the three specific sensitivity experiments, the "population downscaling" (FM20H-P) experiment exhibits the largest performance deviations from their common base experiment (FM20H), followed by the "no biogenic emissions" (FM20H-N) and "coarse meteorology" (FM20H-C) experiments. FM20H-P performs particularly bad for maximum and hourly NO_2 concentrations at industry and traffic sides (see panels g, h, j and k) and particularly well for minimum O_3 concentrations at traffic sides (see panel l). Curiously, among all considered experiments, FM20H-N yields the best results for minimum $PM2.5$ concentrations at industry and traffic sides (see panels i and l) and for maximum NO_2 concentrations at traffic sides (see panel j). The good skill scores at these stations types arise from error compensation effects. Namely, the positive bias is for minimum $PM2.5$, which is smaller at traffic and industry sides than at background sides because the observed concentrations are higher there, is *improved* when biogenic emissions are turned off, which translates into better MAESS values. For maximum NO_2, removing this kind of emissions enhances the temporal correlation, brings the standard deviation closer to unity and finally also improves the MAESS. This, in turn, means that the *inclusion* of biogenic emissions in the remaining experiments deteriorates the temporal variability and day-to-day sequence of the modelled minimum NO_2 time series if compared with observations. At background sides, however, the NO_2 and $PM2.5$ maxima are generally underestimated by the model and the exclusion of biogenic emissions further increases this negative bias (see Fig. 10d and Supplementary Figure 2). The pronounced reduction of the O_3 maxima at background sides in the FM20H-N experiment, as compared with FM20H, points to an active role of biogenic VOCs in this case (see Supplementary Figure 2e). For FM20H-C, deviations from the base experiment are largest for the minima at industry sides and are otherwise generally weak (see Fig. 10i).

4 Discussion and Conclusions

In this study, a series of 19 sensitivity experiments was carried out with the chemical weather forecasting model CHIMERE over the northwestern Iberian Peninsula for the 2018 summer season in order to assess the model's capability to reproduce in-situ NO_2, O_3, $PM10$ and $PM2.5$ surface concentrations on daily to hourly timescale. The range of applied model experiments covers the effects of distinct emission configurations, horizontal and vertical resolution setups and model chemistries. With the help of three secondary experiments, the impact of population downscaling, coarse resolution meteorological input data and missing biogenic emissions is discussed in addition. All these experiments were driven by meteorological data from WRF and chemical boundary data from ECMWF's C-IFS.

The obtained results are very heterogeneous and the applied model improvement efforts, often associated with considerable computational costs, do generally *not* lead to an unrestricted model improvement. For most efforts, verification results im-

prove for *some* aspects but worsen for others. Nonetheless, one single factor has been identified that improves the model in a systematic way, returning better results for virtually all aspects of the verification.

The first take-home message is that the use of an up-to-date and sophistically downscaled anthropogenic emission inventory (configuration 1: EMEP for the year 2017 downscaled with land use, population and traffic proxies as well as large point sources), if compared to an older inventory downscaled with basic options (configuration 2: HTAP v2.2 for the year 2010 downscaled with land use only), on the one hand improves the modelled O_3 and $PM10$ concentrations, but on the other hand deteriorates the results for NO_2 and $PM10$. This is in line with Russo et al. (2019), in the sense that an upgraded emission inventory does not necessarily improve the modelled pollutant concentrations with respect to observations in all aspects.

Second, heterogeneous results are obtained for the performance changes associated with the *chemical mechanism*. While the performance for NO_2 is practically unrelated to the chosen mechanism, the full Melchior mechanism is preferable to SAPRC if O_3 concentrations —at any temporal scale— are considered. For particulate matter, SAPRC is preferable for the daily maxima and hourly concentrations and full Melchior for the daily minima.

Third, an increase in the *horizontal* resolution of the CHIMERE domain and associated emissions from $0.15° \times 0.15°$ to $0.05° \times 0.04°$ does *not* lead to a systematic model improvement but rather to a large increase in the spatial variability of the results. In line with Valari and Menut (2008), we have indications that this is caused by the noise increase in high resolution *meteorological* input data and, to an even larger degree, by the population downscaling procedure used to reallocate the raw data from the applied anthropogenic emission inventories on the subgrid scale. If this kind of downscaling is used, the model overestimates the temporal mean value of the daily maximum and hourly concentrations at traffic and industry sides. The same applies to the temporal standard deviation, i.e. to the model's capability do simulate the degree of temporal variability from one day to another.

Contrary to the effects obtained with an increased horizontal resolution, the use of 20 instead of 10 vertical layers within the lower to middle troposphere (999 to 500 hPa) *systematically improves* the model results for nearly all aspects of the verification.

All together, and as long as top-down emission inventories coming on a relatively coarse spatial and temporal resolution are applied, we recommend the use of 20 model layers together with a horizontal resolution not much finer than the native resolution of the inventory. In this context, the resolution of the coarse domain applied here ($0.15° \times 0.15°$) may not be optimal and in future studies should be approximated to the native grid of the emission inventory (i.e. $0.1° \times 0.1°$ for both HTAP and EMEP) in order to see whether the results can be further improved. Likewise, a *region-specific* optimization of the downscaling procedures used to re-allocate raw emissions on the subgrid scale according to proxy data for population and traffic density would likely yield better results for the northwestern Iberian Peninsula, particularly in what concerns the NO_2 and $PM2.5$ concentrations.

As a final remark, the present study has explored a *broad* range of model performance factors with *empirical* methods, mainly to provide practical recommendations for the numerical modelling community. In the future, our results should be complemented by analytic in-depth studies focussing on single factors.

Code availability. The CHIMERE v2017r4 release is freely available and provided under the GNU general public license. The source code of this model version can be obtained from the CHIMERE web site at https://www.lmd.polytechnique.fr/chimere and is explained in Mailler et al. (2017). The WRF v3.5.1 source code is available from GitHub at https://github.com/wrf-model/WRF and can be also obtained from https://www2.mmm.ucar.edu/wrf/users. The official DOI for WRF-ARW is https://doi.org/10.5065/D6MK6B4K (WRF-Community, 2013) and a reference article about model version 3 was published by Skamarock et al. (2008). Since these source codes are permanently saved on their respective official repositories, there is no need for additional archiving. The configuration files of all CHIMERE and WRF experiments run for the present study have been made permanently available at https://doi.org/10.5281/zenodo.3909451 (Brands, 2020).

Data availability. The CHIMERE model output generated in this study and the observational data used as reference for model verification are available from https://doi.org/10.5281/zenodo.3909451 (Brands, 2020). GFS and C-IFS data are available from https://doi.org/10.5065/D65D8PWK (NCEP et al., 2015) and https://doi.org/10.5194/gmd-8-975-2015 (Flemming et al., 2015). The HTAP v2.2 and EMEP 2017 emission inventories were retrieved from https:doi.org/10.5194/acp-15-11411-2015 (Janssens-Maenhout et al., 2015) and https://www.ceip.at/ms/ceip_home1/ceip_home/new_emep-grid/01_grid_data (EMEP/CEIP, 2019), respectively.

Author contributions. Swen Brands designed and executed the CHIMERE experiments, disaggregated the HTAP emission inventory, built the figures, analyzed the results, wrote the manuscript and supervised the study. Guillermo Fernández-García contributed to the writing of the manuscript and provided WRF data. Marta García disaggregated the EMEP emission inventory and contributed to the writing of the manuscript, Nuria Fernández García, Anthony Estévez Saunders, Paula Costa Tomé and Cristina Otero were responsible for the quality control of the applied observations. Marcos Tesouro Montecelo, Pablo Enrique Carracedo García, Anabela Neto Venâncio and Pedro Daniel Melo Costa provided WRF data. María Luz Macho and Juan Taboada contributed to the writing of the manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. The authors would like to thank the CHIMERE and WRF development teams for providing their source code and technical support. A special thanks goes Laurent Menut and Myrto Valari for their scientific guidance and to Florian Couvedat and Bertrand Bessagnet for sharing their emission downscaling programs. The authors also gratefully acknowledge the computational resources and technical support provided by the Centro de Supercomputación de Galicia (CESGA), as well as the free availability of the global predictions from the GFS and C-IFS forecasting systems maintained by NCEP and ECMWF/Copernicus, respectively.

References

Alfaro, S. C. and Gomes, L.: Modeling mineral aerosol production by wind erosion: Emission intensities and aerosol size distributions in source areas, Journal of Geophysical Research: Atmospheres, 106, 18 075–18 084, https://doi.org/10.1029/2000JD900339, 2001.

Balzarini, A., Pirovano, G., Honzak, L., Žabkar, R., Curci, G., Forkel, R., Hirtl, M., José, R. S., Tuccella, P., and Grell, G.: WRF-Chem model sensitivity to chemical mechanisms choice in reconstructing aerosol optical properties, Atmospheric Environment, 115, 604 – 619, https://doi.org/10.1016/j.atmosenv.2014.12.033, 2015.

Banzhaf, S., Schaap, M., Kerschbaumer, A., Reimer, E., Stern, R., van der Swaluw, E., and Builtjes, P.: Implementation and evaluation of pH-dependent cloud chemistry and wet deposition in the chemical transport model REM-Calgrid, Atmospheric Environment, 49, 378–390, https://doi.org/10.1016/j.atmosenv.2011.10.069, 2012.

Banzhaf, S., Schaap, M., Kranenburg, R., Manders, A. M. M., Segers, A. J., Visschedijk, A. J. H., Denier van der Gon, H. A. C., Kuenen, J. J. P., van Meijgaard, E., van Ulft, L. H., Cofala, J., and Builtjes, P. J. H.: Dynamic model evaluation for secondary inorganic aerosol and its precursors over Europe between 1990 and 2009, Geoscientific Model Development, 8, 1047–1070, https://doi.org/10.5194/gmd-8-1047-2015, 2015.

Beegum, S. N., Gherboudj, I., Chaouch, N., Couvidat, F., Menut, L., and Ghedira, H.: Simulating aerosols over Arabian Peninsula with CHIMERE: Sensitivity to soil, surface parameters and anthropogenic emission inventories, Atmospheric Environment, 128, 185–197, https://doi.org/10.1016/j.atmosenv.2016.01.010, 2016.

Bessagnet, B., Pirovano, G., Mircea, M., Cuvelier, C., Aulinger, A., Calori, G., Ciarelli, G., Manders, A., Stern, R., Tsyro, S., García Vivanco, M., Thunis, P., Pay, M.-T., Colette, A., Couvidat, F., Meleux, F., Rouïl, L., Ung, A., Aksoyoglu, S., Baldasano, J. M., Bieser, J., Briganti, G., Cappelletti, A., D'Isidoro, M., Finardi, S., Kranenburg, R., Silibello, C., Carnevale, C., Aas, W., Dupont, J.-C., Fagerli, H., Gonzalez, L., Menut, L., Prévôt, A. S. H., Roberts, P., and White, L.: Presentation of the EURODELTA III intercomparison exercise – evaluation of the chemistry transport models' performance on criteria pollutants and joint analysis with meteorology, Atmospheric Chemistry and Physics, 16, 12 667–12 701, https://doi.org/10.5194/acp-16-12667-2016, 2016.

Bessagnet, B., Menut, L., Colette, A., Couvidat, F., Dan, M., Mailler, S., Létinois, L., Pont, V., and Rouïl, L.: An evaluation of the CHIMERE Chemistry Transport Model to simulate dust outbreaks across the Northern Hemisphere in March 2014, Atmosphere, 8, https://doi.org/10.3390/atmos8120251, 2017.

Bicheron, P., Amberg, V., Bourg, L., Petit, D., Huc, M., Miras, B., Brockmann, C., Hagolle, O., Delwart, S., Ranera, F., Leroy, M., and Arino, O.: Geolocation Assessment of MERIS GlobCover Orthorectified Products, IEEE Transactions on Geoscience and Remote Sensing, 49, 2972–2982, https://doi.org/10.1109/TGRS.2011.2122337, 2011.

Brands, S.: Underlying experimental data for "An exploratory performance assessment of the CHIMERE model (version 2017r4) for the northwestern Iberian Peninsula and the summer season", https://doi.org/10.5281/zenodo.3909451, 2020.

Brasseur, G. P., Xie, Y., Petersen, A. K., Bouarar, I., Flemming, J., Gauss, M., Jiang, F., Kouznetsov, R., Kranenburg, R., Mijling, B., Peuch, V.-H., Pommier, M., Segers, A., Sofiev, M., Timmermans, R., van der A, R., Walters, S., Xu, J., and Zhou, G.: Ensemble forecasts of air quality in eastern China – Part 1: Model description and implementation of the MarcoPolo–Panda prediction system, version 1, Geoscientific Model Development, 12, 33–67, https://doi.org/10.5194/gmd-12-33-2019, 2019.

Caplan, P., Derber, J., Gemmill, W., Hong, S.-Y., Pan, H.-L., and Parrish, D.: Changes to the 1995 NCEP Operational Medium-Range Forecast Model Analysis–Forecast System, Weather and Forecasting, 12, 581–594, https://doi.org/10.1175/1520-0434(1997)012<0581:CTTNOM>2.0.CO;2, 1997.

Carter, W. P.: Development of the SAPRC-07 chemical mechanism, Atmospheric Environment, 44, 5324–5335, https://doi.org/10.1016/j.atmosenv.2010.01.026, 2010.

Chang, J. C. and Hanna, S. R.: Air quality model performance evaluation, Meteorology and Atmospheric Physics, 87, 167–196, https://doi.org/10.1007/s00703-003-0070-7, 2004.

CLRTAP: Transboundary particulate matter, photo-oxidants, acidifying and eutrophying components, Tech. Rep. Status Report 1/2019, CLRTAP, 2019.

Colette, A., Granier, C., Hodnebrog, Ø., Jakobs, H., Maurizi, A., Nyiri, A., Bessagnet, B., D'Angiola, A., D'Isidoro, M., Gauss, M., Meleux, F., Memmesheimer, M., Mieville, A., Rouïl, L., Russo, F., Solberg, S., Stordal, F., and Tampieri, F.: Air quality trends in Europe over the past decade: a first multi-model assessment, Atmospheric Chemistry and Physics, 11, 11 657–11 678, https://doi.org/10.5194/acp-11-11657-2011, 2011.

Colette, A., Andersson, C., Manders, A., Mar, K., Mircea, M., Pay, M.-T., Raffort, V., Tsyro, S., Cuvelier, C., Adani, M., Bessagnet, B., Bergström, R., Briganti, G., Butler, T., Cappelletti, A., Couvidat, F., D'Isidoro, M., Doumbia, T., Fagerli, H., Granier, C., Heyes, C., Klimont, Z., Ojha, N., Otero, N., Schaap, M., Sindelarova, K., Stegehuis, A. I., Roustan, Y., Vautard, R., van Meijgaard, E., Vivanco, M. G., and Wind, P.: EURODELTA-Trends, a multi-model experiment of air quality hindcast in Europe over 1990–2010, Geoscientific Model Development, 10, 3255–3276, https://doi.org/10.5194/gmd-10-3255-2017, 2017.

Crippa, P., Sullivan, R. C., Thota, A., and Pryor, S. C.: The impact of resolution on meteorological, chemical and aerosol properties in regional simulations with WRF-Chem, Atmospheric Chemistry and Physics, 17, 1511–1528, https://doi.org/10.5194/acp-17-1511-2017, 2017.

Davies, T. D., Kelly, P. M., Low, P. S., and Pierce, C. E.: Surface ozone concentrations in Europe: Links with the regional-scale atmospheric circulation, Journal of Geophysical Research: Atmospheres, 97, 9819–9832, https://doi.org/10.1029/92JD00419, 1992.

Deroubaix, A., Menut, L., Flamant, C., Brito, J., Denjean, C., Dreiling, V., Fink, A., Jambert, C., Kalthoff, N., Knippertz, P., Ladkin, R., Mailler, S., Maranan, M., Pacifico, F., Piguet, B., Siour, G., and Turquety, S.: Diurnal cycle of coastal anthropogenic pollutant transport over southern West Africa during the DACCIWA campaign, Atmospheric Chemistry and Physics, 19, 473–497, https://doi.org/10.5194/acp-19-473-2019, 2019.

EMEP/CEIP: Spatially distributed emission data as used in EMEP models, Tech. rep., Centre on Emission Inventories and Projections, https://www.ceip.at/ms/ceip_home1/ceip_home/new_emep-grid/01_grid_data, 2019.

EU: Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe, Official Journal of the European Union, 2008.

Fécan, F., Marticorena, B., and Bergametti, G.: Parametrization of the increase of the aeolian erosion threshold wind friction velocity due to soil moisture for arid and semiarid areas, Annales Geophysicae, 17, 149–157, https://doi.org/10.1007/s00585-999-0149-7, 1998.

Flamant, C., Deroubaix, A., Chazette, P., Brito, J., Gaetani, M., Knippertz, P., Fink, A. H., de Coetlogon, G., Menut, L., Colomb, A., Denjean, C., Meynadier, R., Rosenberg, P., Dupuy, R., Dominutti, P., Duplissy, J., Bourrianne, T., Schwarzenboeck, A., Ramonet, M., and Totems, J.: Aerosol distribution in the northern Gulf of Guinea: local anthropogenic sources, long-range transport, and the role of coastal shallow circulations, Atmospheric Chemistry and Physics, 18, 12 363–12 389, https://doi.org/10.5194/acp-18-12363-2018, 2018.

Flemming, J., Huijnen, V., Arteta, J., Bechtold, P., Beljaars, A., Blechschmidt, A.-M., Diamantakis, M., Engelen, R. J., Gaudel, A., Inness, A., Jones, L., Josse, B., Katragkou, E., Marecal, V., Peuch, V.-H., Richter, A., Schultz, M. G., Stein, O., and Tsikerdekis, A.: Tropospheric chemistry in the Integrated Forecasting System of ECMWF, Geoscientific Model Development, 8, 975–1003, https://doi.org/10.5194/gmd-8-975-2015, 2015.

Gallego, F. J.: A population density grid of the European Union, Population and Environment, 31, 460–473, https://doi.org/10.1007/s11111-010-0108-y, 2010.

Gama, C., Pio, C., Monteiro, A., Russo, M., Fernandes, A. P., Borrego, C., Baldasano, J. M., and Tchepel, O.: Comparison of Methodologies for Assessing Desert Dust Contribution to Regional PM10 and PM2.5 Levels: A One-Year Study Over Portugal, Atmosphere, 11, 134, https://doi.org/10.3390/atmos11020134, 2020.

Giorgi, F. and Francisco, R.: Uncertainties in regional climate change prediction: a regional analysis of ensemble simulations with the HADCM2 coupled AOGCM, Climate Dynamics, 16, 169–182, https://doi.org/10.1007/PL00013733, 2000.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C.: Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmospheric Chemistry and Physics, 6, 3181–3210, https://doi.org/10.5194/acp-6-3181-2006, 2006.

Im, U., Brandt, J., Geels, C., Hansen, K. M., Christensen, J. H., Andersen, M. S., Solazzo, E., Kioutsioukis, I., Alyuz, U., Balzarini, A., Baro, R., Bellasio, R., Bianconi, R., Bieser, J., Colette, A., Curci, G., Farrow, A., Flemming, J., Fraser, A., Jimenez-Guerrero, P., Kitwiroon, N., Liang, C.-K., Nopmongcol, U., Pirovano, G., Pozzoli, L., Prank, M., Rose, R., Sokhi, R., Tuccella, P., Unal, A., Vivanco, M. G., West, J., Yarwood, G., Hogrefe, C., and Galmarini, S.: Assessment and economic valuation of air pollution impacts on human health over Europe and the United States as calculated by a multi-model ensemble in the framework of AQMEII3, Atmospheric Chemistry and Physics, 18, 5967–5989, https://doi.org/10.5194/acp-18-5967-2018, 2018a.

Im, U., Christensen, J. H., Geels, C., Hansen, K. M., Brandt, J., Solazzo, E., Alyuz, U., Balzarini, A., Baro, R., Bellasio, R., Bianconi, R., Bieser, J., Colette, A., Curci, G., Farrow, A., Flemming, J., Fraser, A., Jimenez-Guerrero, P., Kitwiroon, N., Liu, P., Nopmongcol, U., Palacios-Peña, L., Pirovano, G., Pozzoli, L., Prank, M., Rose, R., Sokhi, R., Tuccella, P., Unal, A., Vivanco, M. G., Yarwood, G., Hogrefe, C., and Galmarini, S.: Influence of anthropogenic emissions and boundary conditions on multi-model simulations of major air pollutants over Europe and North America in the framework of AQMEII3, Atmospheric Chemistry and Physics, 18, 8929–8952, https://doi.org/10.5194/acp-18-8929-2018, 2018b.

Inness, A., Baier, F., Benedetti, A., Bouarar, I., Chabrillat, S., Clark, H., Clerbaux, C., Coheur, P., Engelen, R. J., Errera, Q., Flemming, J., George, M., Granier, C., Hadji-Lazaro, J., Huijnen, V., Hurtmans, D., Jones, L., Kaiser, J. W., Kapsomenakis, J., Lefever, K., Leitão, J., Razinger, M., Richter, A., Schultz, M. G., Simmons, A. J., Suttie, M., Stein, O., Thépaut, J.-N., Thouret, V., Vrekoussis, M., Zerefos, C., and the MACC team: The MACC reanalysis: an 8 yr data set of atmospheric composition, Atmospheric Chemistry and Physics, 13, 4073–4109, https://doi.org/10.5194/acp-13-4073-2013, 2013.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Dentener, F., Muntean, M., Pouliot, G., Keating, T., Zhang, Q., Kurokawa, J., Wankmüller, R., Denier van der Gon, H., Kuenen, J. J. P., Klimont, Z., Frost, G., Darras, S., Koffi, B., and Li, M.: HTAP v2.2: a mosaic of regional and global emission grid maps for 2008 and 2010 to study hemispheric transport of air pollution, Atmospheric Chemistry and Physics, 15, 11 411–11 432, https://doi.org/10.5194/acp-15-11411-2015, 2015.

Jolliffe, I. T. and Stephenson, D. B., eds.: Forecast verification: a practitioner's guide in atmospheric science, Wiley, 2012.

Jonson, J. E., Simpson, D., Fagerli, H., and Solberg, S.: Can we explain the trends in European ozone levels?, Atmospheric Chemistry and Physics, 6, 51–66, https://doi.org/10.5194/acp-6-51-2006, 2006.

Karlický, J., Huszár, P., and Halenka, T.: Validation of gas phase chemistry in the WRF-Chem model over Europe, Advances in Science and Research, 14, 181–186, https://doi.org/10.5194/asr-14-181-2017, 2017.

Kukkonen, J., Olsson, T., Schultz, D. M., Baklanov, A., Klein, T., Miranda, A. I., Monteiro, A., Hirtl, M., Tarvainen, V., Boy, M., Peuch, V.-H., Poupkou, A., Kioutsioukis, I., Finardi, S., Sofiev, M., Sokhi, R., Lehtinen, K. E. J., Karatzas, K., San José, R., Astitha, M., Kallos,

G., Schaap, M., Reimer, E., Jakobs, H., and Eben, K.: A review of operational, regional-scale, chemical weather forecasting models in Europe, Atmospheric Chemistry and Physics, 12, 1–87, https://doi.org/10.5194/acp-12-1-2012, 2012.

Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L., and Merchant, J. W.: Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data, International Journal of Remote Sensing, 21, 1303–1330, https://doi.org/10.1080/014311600210191, 2000.

Mailler, S., Menut, L., Khvorostyanov, D., Valari, M., Couvidat, F., Siour, G., Turquety, S., Briant, R., Tuccella, P., Bessagnet, B., Colette, A., Létinois, L., Markakis, K., and Meleux, F.: CHIMERE-2017: from urban to hemispheric chemistry-transport modeling, Geoscientific Model Development, 10, 2397–2423, https://doi.org/10.5194/gmd-10-2397-2017, 2017.

Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Denier van der Gon, H. A. C., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., Eskes, H. J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadygrov, N., Kaiser, J. W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouïl, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, A., van Velthoven, P., van Versendaal, R., Vira, J., and Ung, A.: A regional air quality forecasting system over Europe: the MACC-II daily ensemble production, Geoscientific Model Development, 8, 2777–2813, https://doi.org/10.5194/gmd-8-2777-2015, 2015.

Markakis, K., Valari, M., Perrussel, O., Sanchez, O., and Honore, C.: Climate-forced air-quality modeling at the urban scale: sensitivity to model resolution, emissions and meteorology, Atmospheric Chemistry and Physics, 15, 7703–7723, https://doi.org/10.5194/acp-15-7703-2015, 2015.

Mazzeo, A., Huneeus, N., Ordoñez, C., Orfanoz-Cheuquelaf, A., Menut, L., Mailler, S., Valari, M., van der Gon, H. D., Gallardo, L., Muñoz, R., Donoso, R., Galleguillos, M., Osses, M., and Tolvett, S.: Impact of residential combustion and transport emissions on air pollution in Santiago during winter, Atmospheric Environment, 190, 195 – 208, https://doi.org/10.1016/j.atmosenv.2018.06.043, 2018.

Menut, L.: Sensitivity of hourly Saharan dust emissions to NCEP and ECMWF modeled wind speed, Journal of Geophysical Research: Atmospheres, 113, D16 201, https://doi.org/10.1029/2007JD009522, 2008.

Menut, L., Schmechtig, C., and Marticorena, B.: Sensitivity of the sandblasting flux calculations to the soil size distribution accuracy, Journal of Atmospheric and Oceanic Technology, 22, 1875–1884, https://doi.org/10.1175/JTECH1825.1, 2005.

Menut, L., Goussebaile, A., Bessagnet, B., Khvorostiyanov, D., and Ung, A.: Impact of realistic hourly emissions profiles on air pollutants concentrations modelled with CHIMERE, Atmospheric Environment, 49, 233–244, https://doi.org/10.1016/j.atmosenv.2011.11.057, 2012.

Menut, L., Bessagnet, B., Colette, A., and Khvorostiyanov, D.: On the impact of the vertical resolution on chemistry-transport modelling, Atmospheric Environment, 67, 370–384, https://doi.org/10.1016/j.atmosenv.2012.11.026, 2013a.

Menut, L., Bessagnet, B., Khvorostyanov, D., Beekmann, M., Blond, N., Colette, A., Coll, I., Curci, G., Foret, G., Hodzic, A., Mailler, S., Meleux, F., Monge, J.-L., Pison, I., Siour, G., Turquety, S., Valari, M., Vautard, R., and Vivanco, M. G.: CHIMERE 2013: a model for regional atmospheric composition modelling, Geoscientific Model Development, 6, 981–1028, https://doi.org/10.5194/gmd-6-981-2013, 2013b.

Menut, L., Flamant, C., Turquety, S., Deroubaix, A., Chazette, P., and Meynadier, R.: Impact of biomass burning on pollutant surface concentrations in megacities of the Gulf of Guinea, Atmospheric Chemistry and Physics, 18, 2687–2707, https://doi.org/10.5194/acp-18-2687-2018, 2018.

Misenis, C. and Zhang, Y.: An examination of sensitivity of WRF/Chem predictions to physical parameterizations, horizontal grid spacing, and nesting options, Atmospheric Research, 97, 315–334, https://doi.org/10.1016/j.atmosres.2010.04.005, 2010.

Monteiro, A., Russo, M., Gama, C., and Borrego, C.: How important are maritime emissions for the air quality: At European and national scale, Environmental Pollution, 242, 565–575, https://doi.org/10.1016/j.envpol.2018.07.011, 2018.

NCEP, NWS, NOAA, and DOC: NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive, https://doi.org/10.5065/D65D8PW, 2015.

O'Hare, G. and Wilby, R.: A Review of Ozone Pollution in the United Kingdom and Ireland with an Analysis Using Lamb Weather Types, The Geographical Journal, 161, 1–20, 1995.

Otero, N., Sillmann, J., Mar, K. A., Rust, H. W., Solberg, S., Andersson, C., Engardt, M., Bergström, R., Bessagnet, B., Colette, A., Couvidat, F., Cuvelier, C., Tsyro, S., Fagerli, H., Schaap, M., Manders, A., Mircea, M., Briganti, G., Cappelletti, A., Adani, M., D'Isidoro, M., Pay, M.-T., Theobald, M., Vivanco, M. G., Wind, P., Ojha, N., Raffort, V., and Butler, T.: A multi-model comparison of meteorological drivers of surface ozone over Europe, Atmospheric Chemistry and Physics, 18, 12 269–12 288, https://doi.org/10.5194/acp-18-12269-2018, 2018.

Palmer, T. N., Doblas-Reyes, F. J., Weisheimer, A., and Rodwell, M. J.: Toward seamless prediction: calibration of climate change projections using seasonal forecasts, Bulletin of the American Meteorological Society, 89, 459–470, https://doi.org/10.1175/BAMS-89-4-459, 2008.

Russo, M. A., Gama, C., and Monteiro, A.: How does upgrading an emissions inventory affect air quality simulations?, Air Quality, Atmosphere & Health, 12, 731–741, https://doi.org/10.1007/s11869-019-00692-x, 2019.

Schaap, M., Cuvelier, C., Hendriks, C., Bessagnet, B., Baldasano, J., Colette, A., Thunis, P., Karam, D., Fagerli, H., Graff, A., Kranenburg, R., Nyiri, A., Pay, M., Rouïl, L., Schulz, M., Simpson, D., Stern, R., Terrenoire, E., and Wind, P.: Performance of European chemistry transport models as function of horizontal resolution, Atmospheric Environment, 112, 90–105, https://doi.org/10.1016/j.atmosenv.2015.04.003, 2015.

Shao, Y. and Lu, H.: A simple expression for wind erosion threshold friction velocity, Journal of Geophysical Research: Atmospheres, 105, 22 437–22 443, https://doi.org/10.1029/2000JD900304, 2000.

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Wang, W., and Powers, J.: A description of the Advanced Research WRF version 3, NCAR Technical Note NCAR/TN–475+STR, National Center for Atmospheric Research, Boulder, Colorado, USA, https://doi.org/10.5065/D68S4MVH, 126pp., 2008.

Teixeira, J., Carvalho, A., Tuccella, P., Curci, G., and Rocha, A.: WRF-chem sensitivity to vertical resolution during a saharan dust event, Physics and Chemistry of the Earth, Parts A/B/C, 94, 188 – 195, https://doi.org/10.1016/j.pce.2015.04.002, 2016.

Terrenoire, E., Bessagnet, B., Rouïl, L., Tognet, F., Pirovano, G., Létinois, L., Beauchamp, M., Colette, A., Thunis, P., Amann, M., and Menut, L.: High-resolution air quality simulation over Europe with the chemistry transport model CHIMERE, Geoscientific Model Development, 8, 21–42, https://doi.org/10.5194/gmd-8-21-2015, 2015.

Theobald, M. R., Vivanco, M. G., Aas, W., Andersson, C., Ciarelli, G., Couvidat, F., Cuvelier, K., Manders, A., Mircea, M., Pay, M.-T., Tsyro, S., Adani, M., Bergström, R., Bessagnet, B., Briganti, G., Cappelletti, A., D'Isidoro, M., Fagerli, H., Mar, K., Otero, N., Raffort, V., Roustan, Y., Schaap, M., Wind, P., and Colette, A.: An evaluation of European nitrogen and sulfur wet deposition and their trends estimated by six chemistry transport models for the period 1990–2010, Atmospheric Chemistry and Physics, 19, 379–405, https://doi.org/10.5194/acp-19-379-2019, 2019.

Thunis, P., Degraeuwe, B., Pisoni, E., Meleux, F., and Clappier, A.: Analyzing the efficiency of short-term air quality plans in European cities, using the CHIMERE air quality model, Air. Qual. Atmos. Health., 10, 235–248, https://doi.org/10.1007/s11869-016-0427-y, 2016.

Valari, M. and Menut, L.: Does an increase in air quality models' resolution bring surface ozone concentrations closer to reality?, Journal of Atmospheric and Oceanic Technology, 25, 1955–1968, https://doi.org/10.1175/2008JTECHA1123.1, 2008.

van Leer, B.: Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method, Journal of Computational Physics, 32, 101–136, https://doi.org/10.1016/0021-9991(79)90145-1, 1979.

Vautard, R., Szopa, S., Beekmann, M., Menut, L., Hauglustaine, D. A., Rouil, L., and Roemer, M.: Are decadal anthropogenic emission reductions in Europe consistent with surface ozone observations?, Geophysical Research Letters, 33, https://doi.org/10.1029/2006GL026080, 2006.

Vivanco, M. G., Theobald, M. R., García-Gómez, H., Garrido, J. L., Prank, M., Aas, W., Adani, M., Alyuz, U., Andersson, C., Bellasio, R., Bessagnet, B., Bianconi, R., Bieser, J., Brandt, J., Briganti, G., Cappelletti, A., Curci, G., Christensen, J. H., Colette, A., Couvidat, F., Cuvelier, C., D'Isidoro, M., Flemming, J., Fraser, A., Geels, C., Hansen, K. M., Hogrefe, C., Im, U., Jorba, O., Kitwiroon, N., Manders, A., Mircea, M., Otero, N., Pay, M.-T., Pozzoli, L., Solazzo, E., Tsyro, S., Unal, A., Wind, P., and Galmarini, S.: Modeled deposition of nitrogen and sulfur in Europe estimated by 14 air quality model systems: evaluation, effects of changes in emissions and implications for habitat protection, Atmospheric Chemistry and Physics, 18, 10 199–10 218, https://doi.org/10.5194/acp-18-10199-2018, 2018.

Wilson, R. C., Fleming, Z. L., Monks, P. S., Clain, G., Henne, S., Konovalov, I. B., Szopa, S., and Menut, L.: Have primary emission reduction measures reduced ozone across Europe? An analysis of European rural background ozone trends 1996–2005, Atmospheric Chemistry and Physics, 12, 437–454, https://doi.org/10.5194/acp-12-437-2012, 2012.

WRF-Community: Weather Research and Forecasting (WRF) Model. Version 3.5.1, UCAR/NCAR, https://doi.org/10.5065/D6MK6B4K, 2013.

Zhang, L., Gong, S., Padro, J., and Barrie, L.: A size-segregated particle dry deposition scheme for an atmospheric aerosol module, Atmospheric Environment, 35, 549–560, https://doi.org/10.1016/S1352-2310(00)00326-5, 2001.

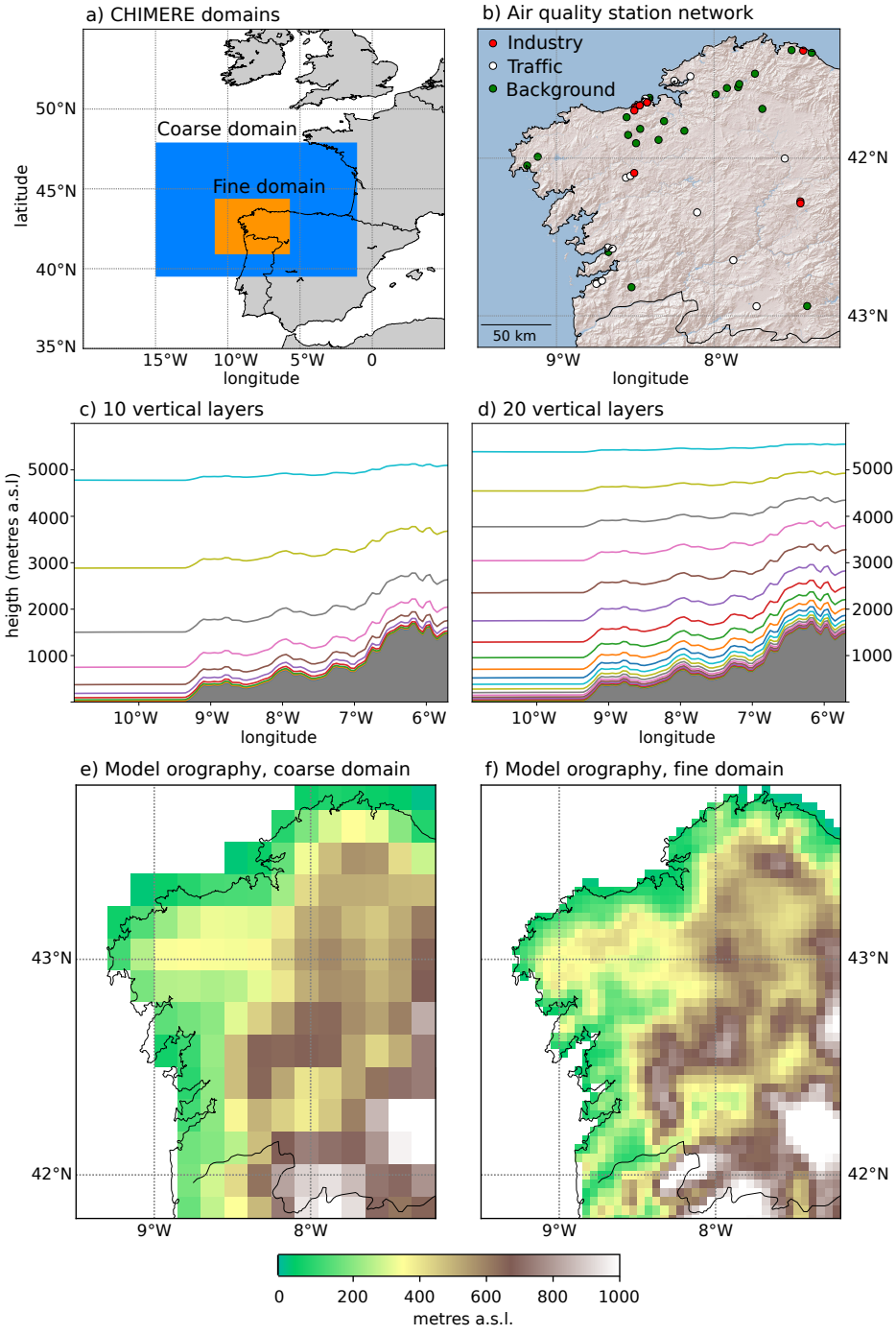

Figure 1. (a) Horizontal CHIMERE domains used for all sensitivity experiments. The fine domain (orange rectangle) is nested into the coarser one (blue rectangle). At the lateral boundary conditions of the coarse domain, CHIMERE is fed by time varying C-IFS data. (b) The Galician air quality station network, grouped by the main pollution sources, (c) Height of the model layers in CHIMERE along 43 °N when using the 10 layer setup and the fine domain, (d) as c but for the 20 layer setup (e) model orography in metres above sea-level (a.s.l.) for the coarse domain, (f) as e but for the fine domain.

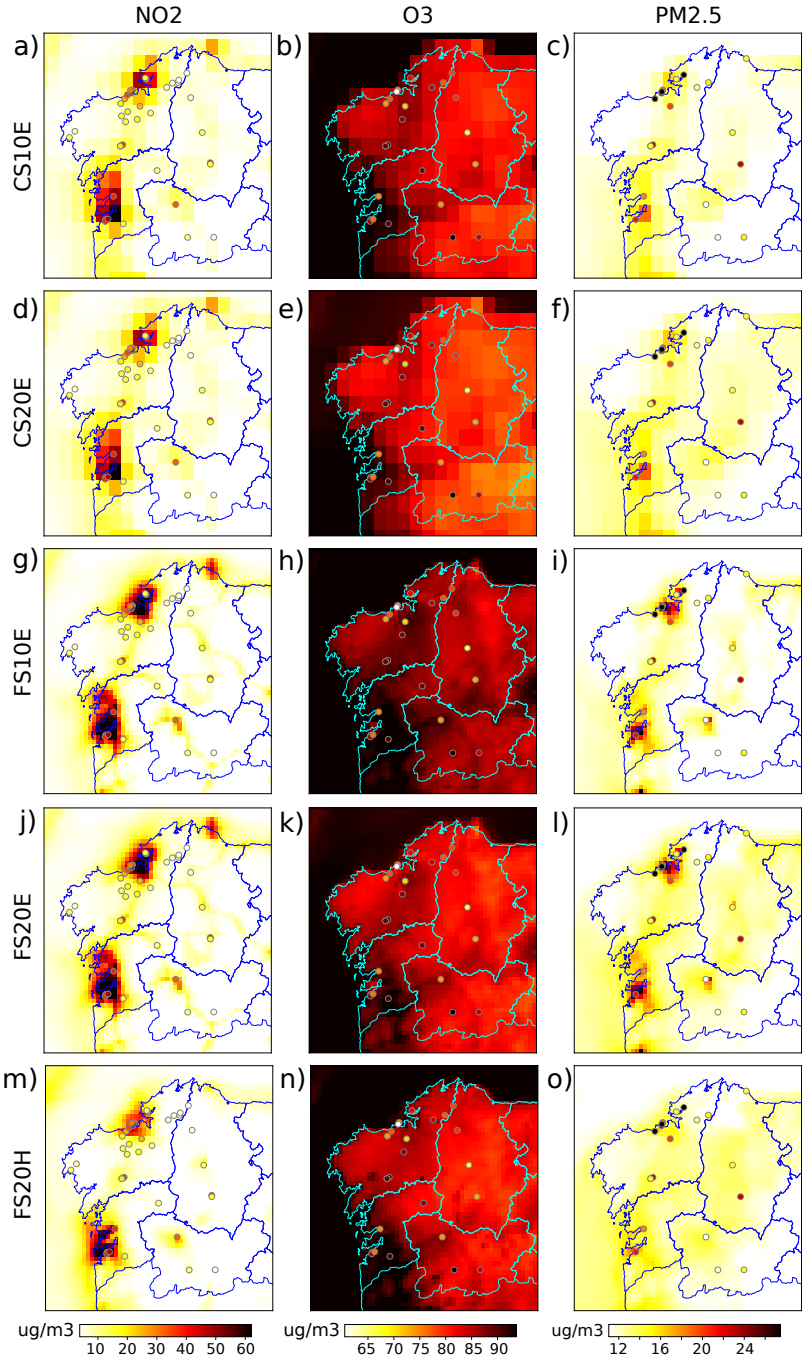

Figure 2. Observed (dots) vs. modelled (underlying pattern) temporal *mean* values for the daily maximum concentrations of NO_2, O_3 and $PM2.5$ and the 5 experiments marked with an asterisk in Table 3, all run with the SAPRC mechanism. The spatial verification results for each panel are provided in Table 4a.

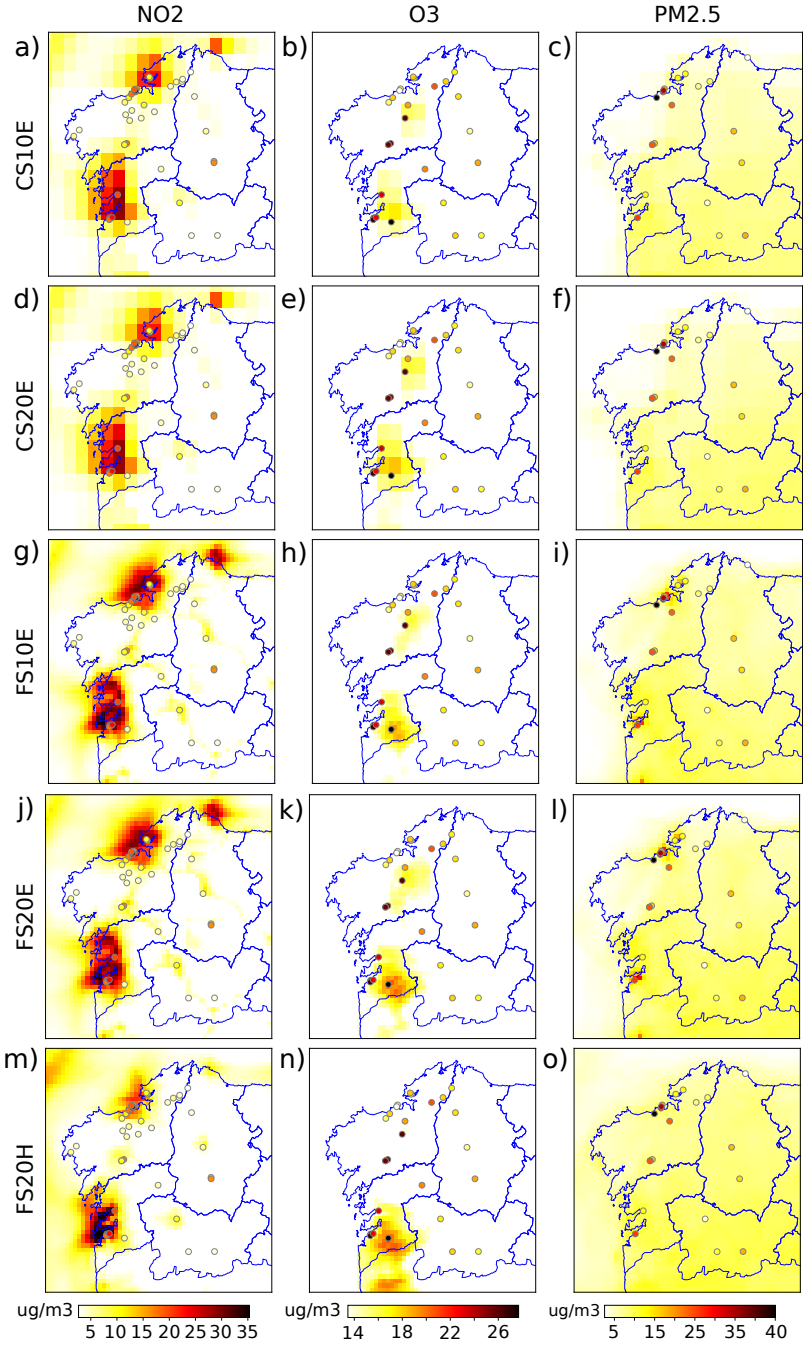

Figure 3. Observed (dots) vs. modelled (underlying pattern) temporal *standard deviation* values for the daily maximum concentrations of NO_2, O_3 and $PM2.5$ and the 5 experiments marked with an asterisk in Table 3), all run with the SAPRC mechanism. The spatial verification results for each panel are provided in Table 4b.

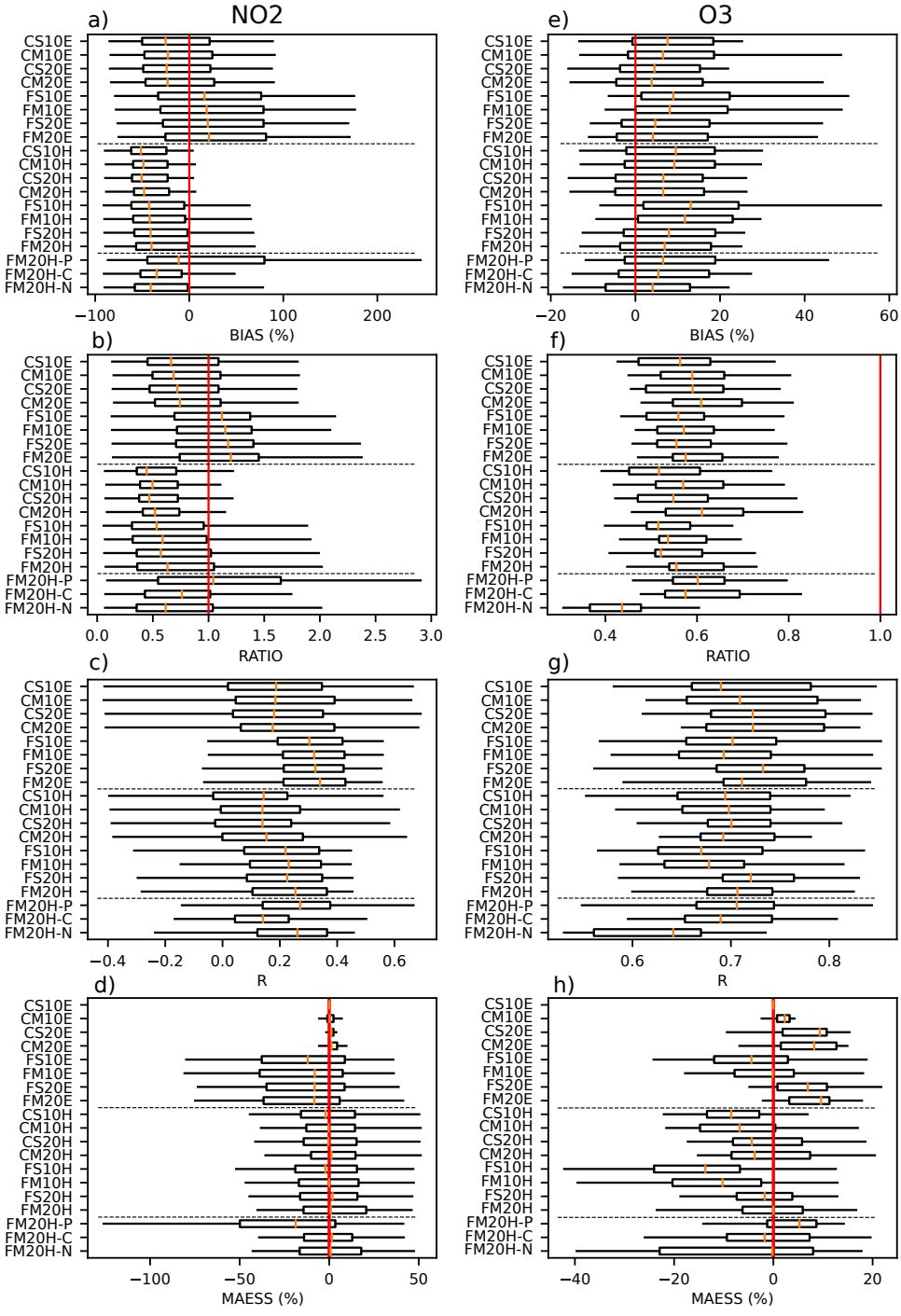

Figure 4. Temporal verification results for daily near-surface *maximum* NO_2 (left) and O_3 (right). Row 1: percentage bias (BIAS), row 2: Pearson correlation coefficient (R), row 3: ratio of standard deviations (RATIO), row 4: mean absolute error skill score (MAESS) with reference to the base experiment CS10E. Boxplots are calculated upon the point-wise verification results at all available stations. Experiments are explained and grouped as in Table 3.

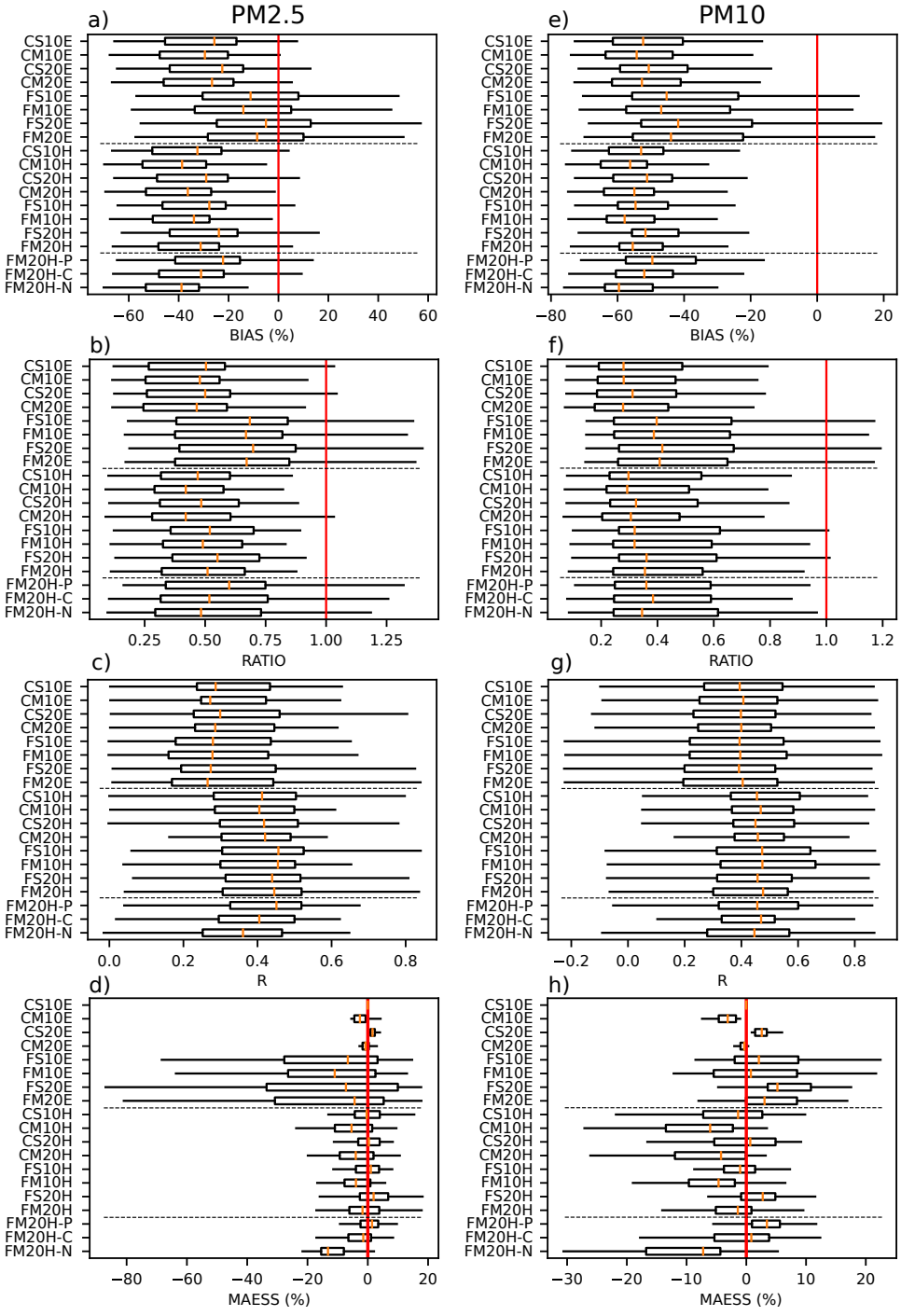

Figure 5. Temporal verification results for daily near-surface *maximum* $PM2.5$ (left) and $PM10$ (right). Row 1: percentage bias (BIAS), row 2: Pearson correlation coefficient (R), row 3: ratio of standard deviations (RATIO), row 4: mean absolute error skill score (MAESS) with reference to the base experiment CS10E. Boxplots are calculated upon the point-wise verification results at all available stations. Experiments are explained and grouped as in Table 3.

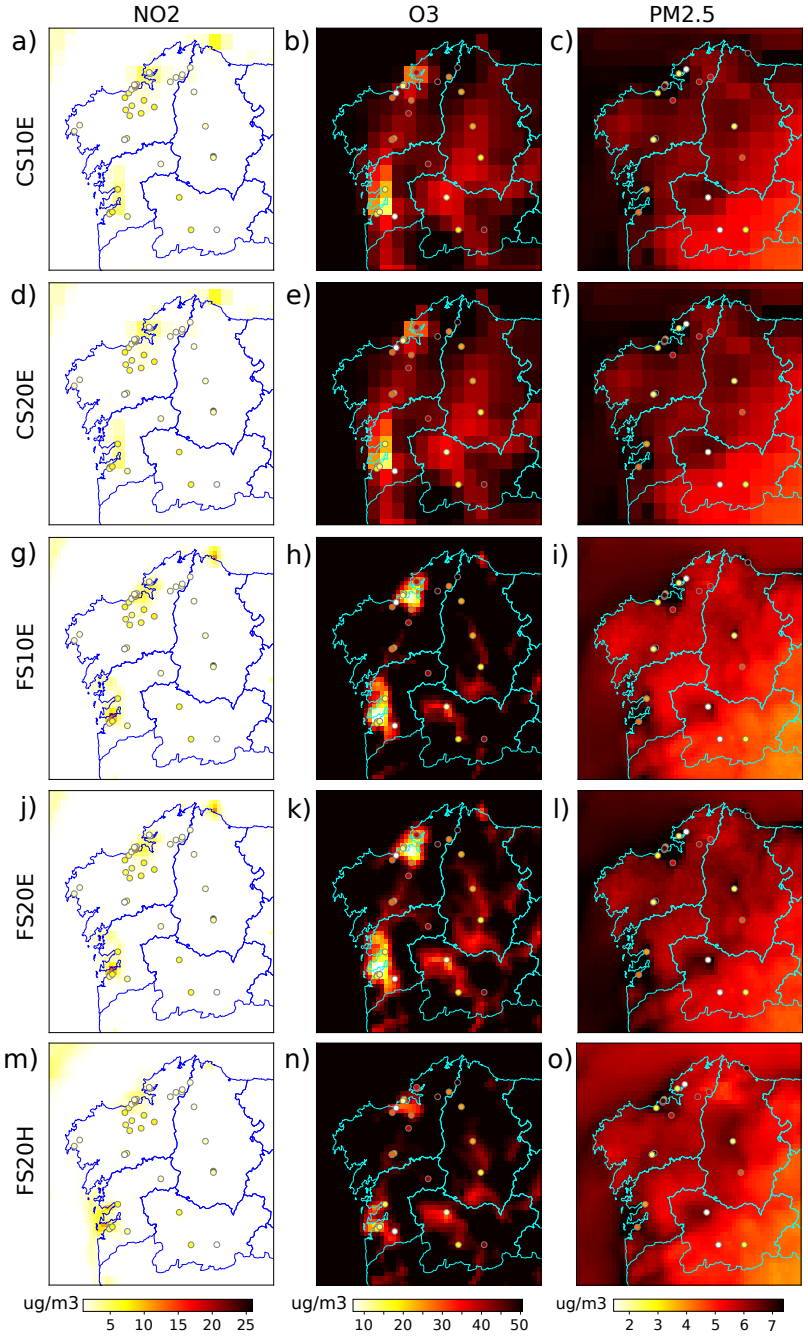

Figure 6. Observed (dots) vs. modelled (underlying pattern) temporal *mean* values for the daily minimum concentrations of NO_2, O_3 and $PM2.5$ and for the five 5 experiments marked with in asterisk in Table 3), all run with the SAPRC mechanism. The spatial verification results for each panel are provided in Table 4c.

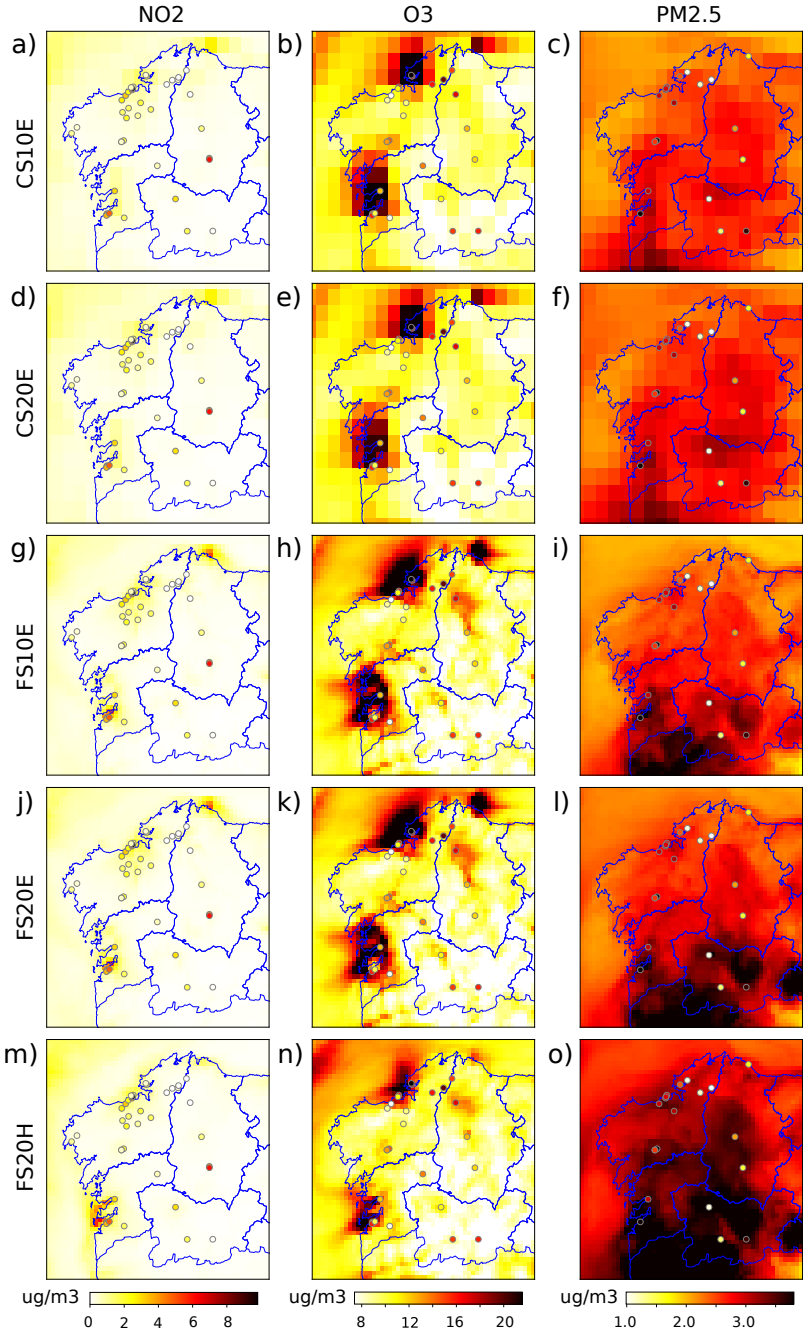

Figure 7. Observed (dots) vs. modelled (underlying pattern) temporal *standard deviation* values for the daily minimum concentrations of NO_2, O_3 and $PM2.5$ and the 5 experiments marked with an asterisk in Table 3, all run with the SAPRC mechanism. The spatial verification results for each panel are provided in Table 4d.

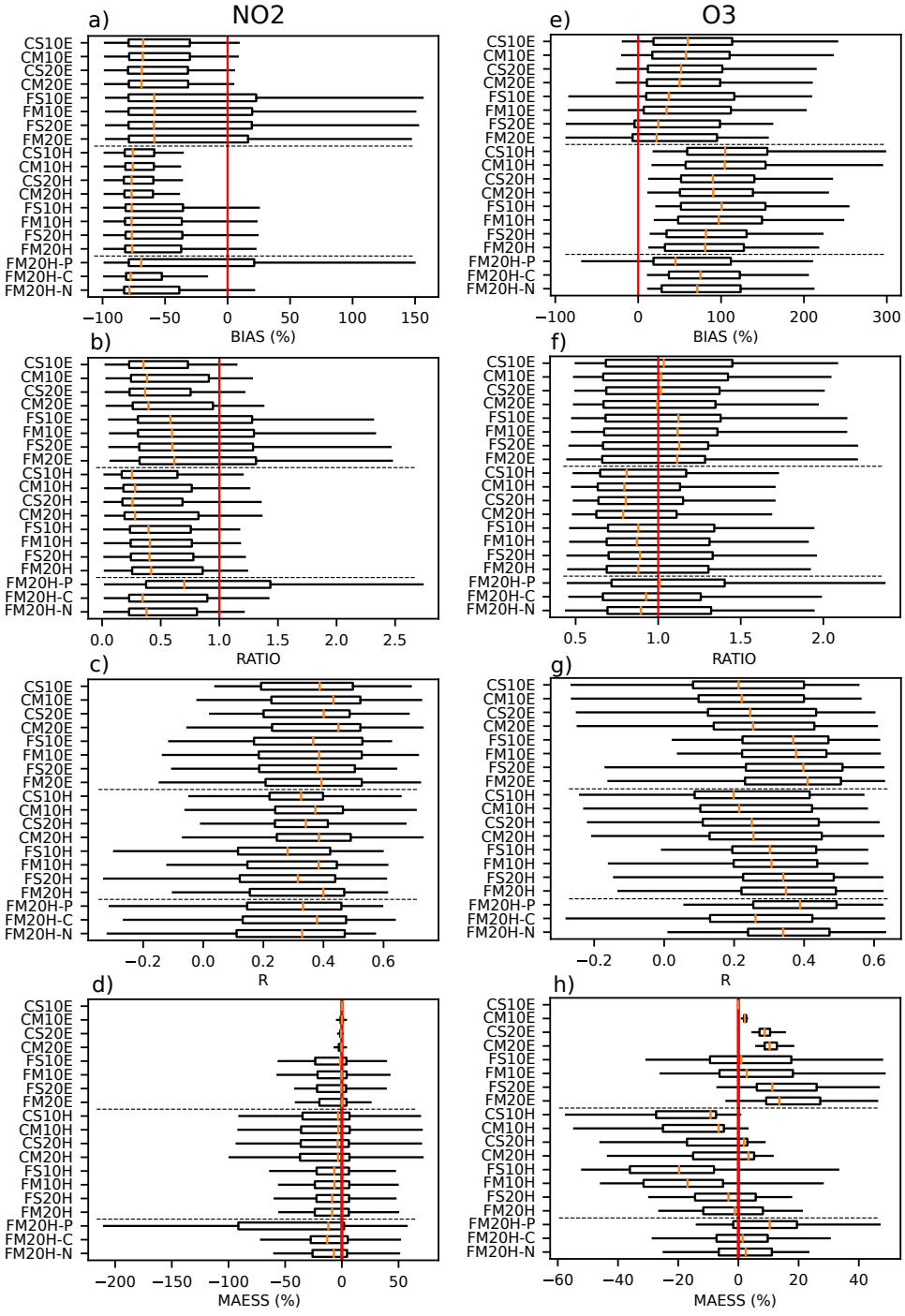

Figure 8. Temporal verification results for daily near-surface *minimum* NO_2 (left) and O_3 (right). Row 1: percentage bias (BIAS), row 2: Pearson correlation coefficient (R), row 3: ratio of standard deviations (RATIO), row 4: mean absolute error skill score (MAESS) with reference to the base experiment CS10E. Boxplots are calculated upon the point-wise verification results at all available stations. Experiments are explained and grouped as in Table 3.

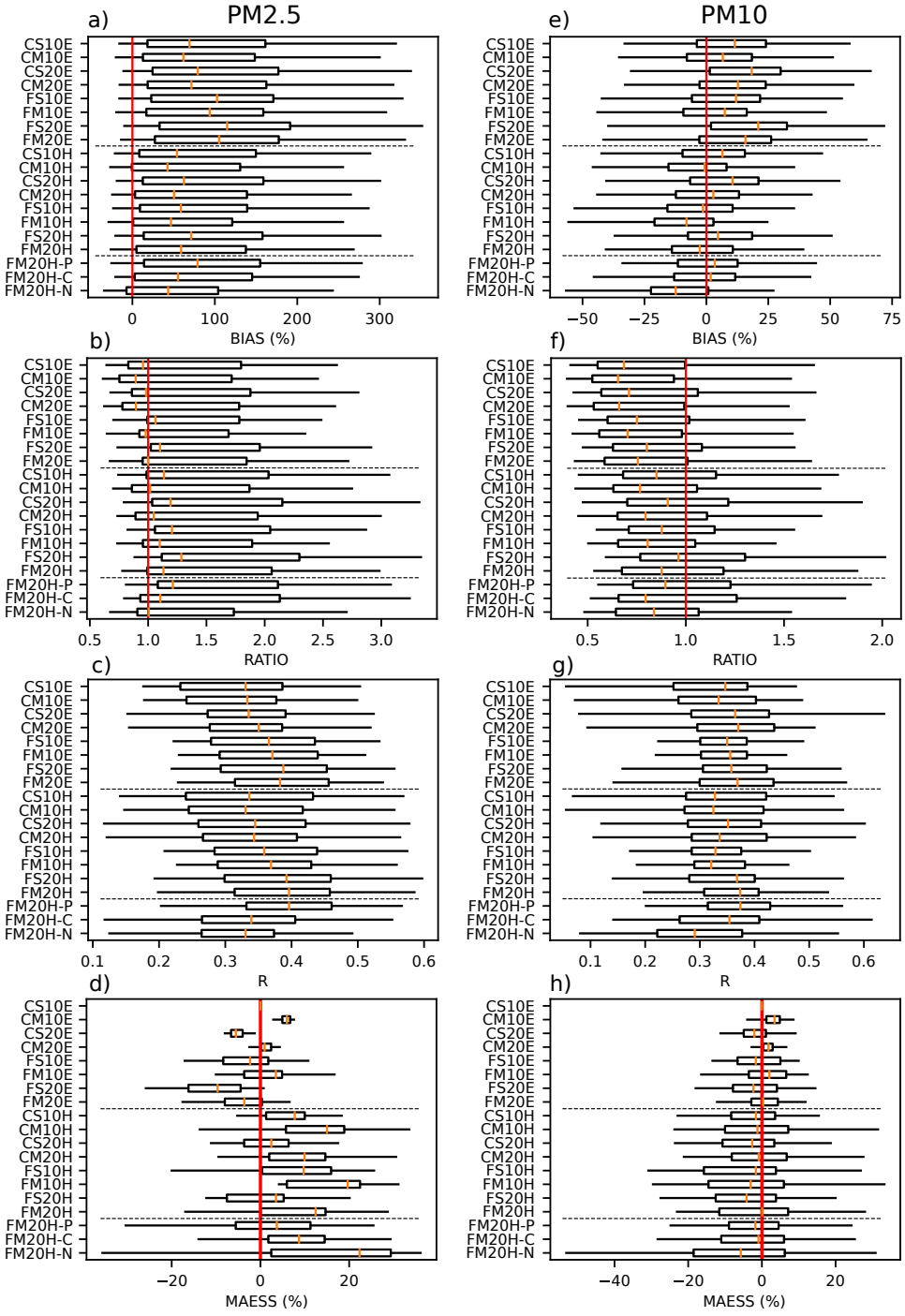

Figure 9. Temporal verification results for daily near-surface *minimum* $PM2.5$ (left) and $PM10$ (right). Row 1: percentage bias (BIAS), row 2: Pearson correlation coefficient (R), row 3: ratio of standard deviations (RATIO), row 4: mean absolute error skill score (MAESS) with reference to the base experiment CS10E. Boxplots are calculated upon the point-wise verification results at all available stations. Experiments are explained and grouped as in Table 3.

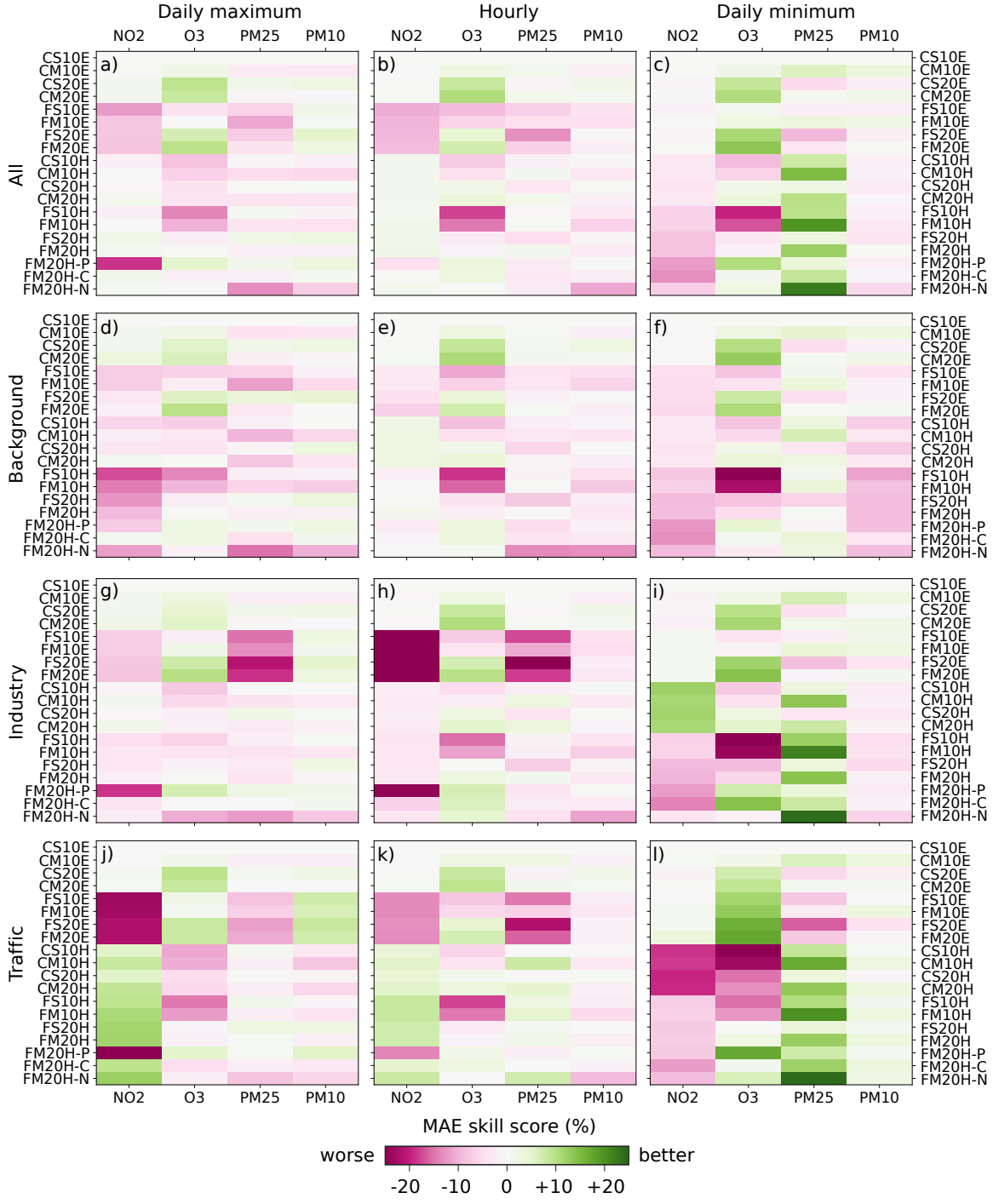

Figure 10. Spatial median mean absolute error skill score (MAESS) with respect to the base experiment CS10E for daily maximum, hourly or daily minimum concentrations (columns 1 to 3 respectively) at all available stations (row 1) or at background, industrial or traffic stations (row 2 to 4 respectively).

Table 1. *WRF* physics common to all sensitivity tests

Parameter	Option
Microphysics	WRF single-moment 6-class scheme
Longwave radiation	Rapid Radiative Transfer Model
Shortwave radiation	Dudhia scheme
Surface layer	MM5 similarity
Land surface	5-layer thermal diffusion
Planetary boundary layer	Yonsei University scheme
Cumulus	Kain-Fritsch scheme

Table 2. *CHIMERE* parameters common to all sensitivity tests

Parameter	Option
Nr. Gauss-Seidel iterations	3
Chemical time-step	adaptive
Physical time-step	5 minutes
Nr. of aerosol size sections	9
Chemically-active aerosols	yes
Sea-salt emission parameterization	inert, parametrization 0
Biogenic emissions	MEGAN
Mineral dust emission	On
Saltation and sandblasting scheme	Alfaro and Gomes (2001), Menut et al. (2005)
Wind threshold estimation	Shao and Lu (2000)
Effect of soil moisture on mineral dust emissions	Fécan et al. (1998)
Secondary organic aerosol scheme	medium complexity
ISORROPIA coupling	yes
Inclusion of carbonaceous species	yes
Aerosol dry deposition	Zhang et al. (2001)
Horizontal advection scheme	van Leer
Vertical advection scheme	upwind
Urban correction	off
Resuspension process	off
Deep convection	on
Lateral boundary conditions	from C-IFS or MACC

Table 3. Overview of the applied sensitivity tests, C = coarse horizontal resolution, F = fine horizontal resolution, 10 = number of vertical layers, S = SAPRC, M = full Melchior, E = EMEP, H = HTAP, P = population downscaling, C = coarse meteorology, N = no biogenic emissions, lu = landuse, popul = population, lsp = emission allocation according to large point sources, Runtime in seconds for a typical summertime heat day (August 5th, 2018)

Acronym	Bio. Emis.	Anth. Emis.	Downscaling	Lu database	Hor. Res. (lat. × lon.)	Layers	Mechanism	Runtime
CS10E*	MEGAN	EMEP	lu, popul, traffic, lsp	GlobCover	WRF: 12×12km, CH: $0.15^\circ \times 0.15^\circ$	10	SAPRC	436s
CM10E	"	"	"	"	"	"	Full Melchior	437s
CS20E*	"	"	"	"	"	20	SAPRC	928s
CM20E	"	"	"	"	"	"	Full Melchior	947s
FS10E*	"	"	"	"	WRF: 4×4km, CH: $0.05^\circ \times 0.04^\circ$	10	SAPRC	1598s
FM10E	"	"	"	"	"	"	Full Melchior	1633s
FS20E*	"	"	"	"	"	20	SAPRC	3582s
FM20E*	"	"	"	"	"	"	Full Melchior	3755s
CS10H	"	HTAP	lu	USGS	WRF: 12×12km, CH: $0.15^\circ \times 0.15^\circ$	10	SAPRC	not saved
CM10H	"	"	"	"	"	"	Full Melchior	"
CS20H	"	"	"	"	"	20	SAPRC	"
CM20H	"	"	"	"	"	"	Full Melchior	"
FS10H	"	"	"	"	WRF: 4×4km, CH: $0.05^\circ \times 0.04^\circ$	10	SAPRC	”
FM10H	"	"	"	"	"	"	Full Melchior	"
FS20H	"	"	"	"	"	20	SAPRC	"
FM20H	"	"	"	"	"	"	Full Melchior	"
FM20H-P	"	"	lu, popul	"	"	"	"	"
FM20H-C	"	"	lu	"	WRF: 12×12km, CH: $0.05^\circ \times 0.04^\circ$	"	"	"
FM20H-N	None	"	"	"	WRF: 4×4km, CH: $0.05^\circ \times 0.04^\circ$	"	"	"

Table 4. Spatial verification results for Figures 2, 3, 6 and 7 are displayed in table section a, b, c and d respectively. Shown is the spatial mean difference (bias, in $\mu g/m^3$), correlation coefficient, standard deviation ratio (modelled/observed) and mean absolute error (in $\mu g/m^3$) calculated upon the modelled and observed temporal mean or standard deviation values at the stations shown in these figures. SBIAS = spatial bias, SR = spatial correlation coefficient, SRATIO = spatial standard deviation ratio, SMAE = spatial mean absolute error, Mean = results for the pointwise temporal mean values, STD = results for the pointwise temporal standard deviation values, Max = results for daily maximum concentrations, Min = results for daily minimum concentrations

Experiment	NO_2				O_3				$PM2.5$			
a) Mean of Max	SBIAS	SR	SRATIO	SMAE	SBIAS	SR	SRATIO	SMAE	SBIAS	SR	SRATIO	SMAE
CS10E	-1.47	0.81	1.39	9.63	6.28	-0.06	0.56	9.38	-6.04	0.35	0.37	6.27
CS20E	-1.50	0.81	1.36	9.42	3.86	-0.05	0.56	8.50	-5.54	0.34	0.36	5.83
FS10E	9.65	0.82	2.14	15.89	8.70	0.03	0.40	9.54	-1.12	0.33	1.28	6.00
FS20E	9.63	0.83	2.08	15.40	5.14	-0.02	0.40	8.00	-0.22	0.30	1.29	6.08
FS20H	-3.02	0.85	1.39	10.39	6.61	-0.10	0.61	9.90	-5.52	0.37	0.27	5.92
b) STD of Max	SBIAS	SR	SRATIO	SMAE	SBIAS	SR	SRATIO	SMAE	SBIAS	SR	SRATIO	SMAE
CS10E	-0.99	0.76	1.21	4.61	-8.82	0.78	0.76	8.82	-6.28	0.00	0.17	6.46
CS20E	-0.94	0.76	1.19	4.51	-8.32	0.77	0.79	8.32	-6.27	0.06	0.19	6.44
FS10E	1.43	0.75	1.22	4.83	-8.88	0.71	0.64	8.86	-3.16	0.48	0.74	4.35
FS20E	1.63	0.75	1.20	4.88	-8.52	0.72	0.69	8.52	-2.89	0.53	0.76	4.22
FS20H	-0.86	0.75	1.45	5.15	-9.07	0.74	0.70	9.07	-5.11	0.37	0.16	5.51
c) Mean of Min	SBIAS	SR	SRATIO	SMAE	SBIAS	SR	SRATIO	SMAE	SBIAS	SR	SRATIO	SMAE
CS10E	-2.93	0.21	0.40	3.41	15.65	0.71	0.97	16.12	2.46	0.28	0.43	2.62
CS20E	-2.97	0.23	0.39	3.43	13.21	0.72	0.93	13.87	2.78	0.30	0.43	2.89
FS10E	-1.80	0.20	0.87	3.81	11.95	0.65	1.48	16.12	2.72	0.17	0.65	3.00
FS20E	-1.87	0.24	0.85	3.78	8.53	0.66	1.38	13.47	3.17	0.12	0.64	3.28
FS20H	-3.07	0.37	0.41	3.52	18.65	0.71	0.87	18.65	2.15	0.08	0.39	2.49
d) STD of Min	SBIAS	SR	SRATIO	SMAE	SBIAS	SR	SRATIO	SMAE	SBIAS	SR	SRATIO	SMAE
CS10E	-1.38	0.36	0.23	1.56	0.39	0.03	1.49	4.96	0.14	0.28	0.28	0.78
CS20E	-1.37	0.35	0.23	1.55	0.03	0.03	1.39	4.65	0.25	0.28	0.28	0.76
FS10E	-0.88	0.43	0.58	1.26	0.79	-0.04	1.53	4.97	0.42	0.48	0.43	0.70
FS20E	-0.86	0.44	0.59	1.25	0.39	-0.04	1.43	4.81	0.59	0.38	0.43	0.77
FS20H	-1.29	0.52	0.37	1.41	-0.26	-0.02	1.25	4.97	0.99	0.36	0.40	1.05